From RNA to histological images: linking the transcriptome with human phenotypes through statistical learning

by

Manuel Muñoz Aguirre

Submitted to the Statistics and Operations Research department in partial fulfillment of the requirements for the degree of Doctor

at the

January 25, 2021

Author .....
Manuel Muñoz Aguirre
Statistics and Operations Research department

Certified by .....
Roderic Guigó Serra
Professor, UPF
Thesis director

Certified by .....
Jan Graffelman
Associate professor, UPC
Thesis co-director

From RNA to histological images: linking the transcriptome with human phenotypes through statistical learning

by

Manuel Muñoz Aguirre

Submitted to the Statistics and Operations Research department
on January 25, 2021, in partial fulfillment of the
requirements for the degree of Doctor

Abstract

Genomic datasets are fundamental to broaden our understanding of human biology in the context of health and disease. However, the high-dimensional nature of gene expression and other molecular traits poses a challenge when attempting to find associations of these data types with human phenotypes. To this end, this thesis relies on statistical learning tools to mitigate the curse of dimensionality and link the human transcriptome with phenotypes at different orders of complexity: from RNA, to computationally-inferred cell type enrichments, and finishing with histological images and their corresponding free-text descriptions. We make four specific contributions. First, we built computational models based on gene expression of post-mortem human tissues in order to derive estimates of post mortem interval. Second, we redefined the basic histological types of tissue classification based on five broad transcriptional programs which define major cell types: epithelial, endothelial, mesenchymal, neural, and blood. We generated computational estimates for the enrichment of these major cell types and validated them through the analysis of histological images and free-text pathology reports, finding that departures from normal cellular enrichment correlate with disease-associated histological phenotypes. Third, we characterized the landscape of human sex-differential gene expression, finding that effects are small but ubiquitous and tend to be tissue-specific, with some of these genes being involved in biological and molecular functions related to disease and clinical phenotypes. Fourth, we proposed an in-silico methodology to spatially deconvolute gene expression from matched sample pairs of whole slide histological images and bulk RNA-seq gene expression, with the goal of replicating the spatial transcriptomics experimental technology. Within this study, we also developed a software tool to effortlessly process whole slide histological images into tiles for machine learning applications.

Thesis director: Roderic Guigó Serra
Title: Professor, UPF

Thesis co-director: Jan Graffelman
Title: Associate professor, UPC

From RNA to histological images: linking the transcriptome with human phenotypes through statistical learning

por

Manuel Muñoz Aguirre

Enviado al departamento de Estadística e Investigación Operativa
el 25 de Enero de 2021, en cumplimiento parcial de
los requisitos para el título de Doctor

Resumen

Los conjuntos de datos genómicos son fundamentales para ampliar nuestra comprensión de la biología humana en el contexto de la salud y la enfermedad. Sin embargo, la alta dimensionalidad de la expresión génica y otros rasgos moleculares constituye un desafío para vincular estos tipos de datos con fenotipos humanos. Esta tesis se apoya en herramientas de aprendizaje estadístico para mitigar el problema de la dimensionalidad y vincular el transcriptoma humano con fenotipos a diferentes niveles de complejidad: desde el ARN, pasando por enriquecimientos de tipos celulares inferidos computacionalmente, y terminando con imágenes histológicas y sus correspondientes anotaciones en formato de texto libre. Hacemos cuatro aportaciones específicas. Primero, construimos modelos computacionales basados en la expresión génica de tejidos humanos post-mortem para realizar estimaciones del intervalo post-mortem. En segundo lugar, redefinimos los tipos histológicos básicos de clasificación de tejidos con base en cinco amplios programas transcripcionales que definen tipos celulares principales: epitelial, endotelial, mesenquimal, neural y sanguíneo. Generamos estimaciones computacionales para el enriquecimiento de estos tipos celulares principales y las validamos mediante el análisis de imágenes histológicas e informes de patología en formato de texto libre, encontrando que las desviaciones respecto a la normalidad en los enriquecimientos celulares correlacionan con fenotipos histológicos asociados con enfermedades. En tercer lugar, caracterizamos el panorama de la expresión diferencial de los genes respecto al sexo en humanos, y descubrimos que los efectos son pequeños pero ubicuos y tienden a ser específicos al tejido, con algunos de estos genes involucrados en funciones biológicas y moleculares relacionadas con enfermedades y fenotipos clínicos. En cuarto lugar, hemos propuesto una metodología in-silico para deconvolucionar espacialmente la expresión génica a partir de muestras emparejadas de imágenes histológicas y expresión génica (bulk RNA-seq), con el objetivo de replicar la tecnología experimental de transcriptómica espacial. Dentro de este estudio, también desarrollamos una herramienta de software para procesar imágenes histológicas y generar mosaicos de imágenes para aplicaciones de aprendizaje automático.

Acknowledgments

I would like to thank my thesis director, Roderic Guigó, for all his feedback, and for providing a research environment that was intellectually stimulating and in which I could sharpen my skills and views on science. My research was also co-supervised by Jan Graffelman, to whom I am also grateful for the feedback and for his guidance in the teaching component of my PhD.

My collaborators have been crucial for the work performed in this thesis. I would like to particularly thank Meritxell Oliva, Sarah Kim-Hellmuth, Barbara Stranger, Pedro Ferreira, Alessandra Breschi, Valentin Wucher, and Vasilis Ntasis. I learned so much by doing research with you. I would also like to thank the GTEx consortium for the repeated opportunities to present my work and for providing a very interesting data resource to work with; as well as Ferran Marques and Verónica Vilaplana from the UPC Image Processing group for useful discussions. The suggestions provided by the Guigó lab members and from my thesis committee (Cedric Notredame and Guillaume Filion) through the years have also been useful to improve the contents of this work. I would also like to thank Romina Garrido, whose administrative support has always been exceptional, and Emilio Palumbo for all the help with computing topics.

The following non-exhaustive list of people shaped the experiences I had in the years of my PhD, and for that, I am thankful:

Ferran Reverter, for your mentorship during the early stages of my time at CRG and help in navigating the learning curve into the field of computational biology, I learned a lot from you about statistics but also about how to discuss results with others.

Meritxell Oliva, it was a lot of fun doing science together, your insights helped realize the importance of trying to see the big picture of what we do.

Raziel Amador, for the fun hackaton times we shared, solving problems that were different to the ones we are used to was very refreshing.

Diego Garrido and Beatrice Borsari, for your friendship and all the adventures we

had through the PhD, it has been quite the journey and I feel glad for having being able to share it with you.

Valentin Wucher, for all the technical, but especially, personal advice through these years. I will always cherish our scientific and meta-scientific conversations, you are truly an example of how fun science can be when done with friends.

Alejandra Arámburo, Luis Abraham Castro, Fernando Campos and Claudia Monge, despite the distance, you have always been a positive presence in my life.

Miguel Rodríguez, for all the spider pranks and jump scares, and from 'saving' me from the worst food-related choking episode ever.

Armando Reyes, for always checking up on me and being someone with whom I can place my undivided trust.

Zara Alaverdyan, for our trips and adventures, listening to my random and oftentimes repetitive nonsense, and simply for being the best.

Mikel Muñoz, I cannot put into words how much you have helped me get to this point. I will always be grateful for that, as well as for your occasional reminders about the things that really matter.

Kaiser Co, for helping me think out stuff so many times, but most importantly, for being my best friend. I am lucky to have you in my life.

Finalmente, y sobre todo, quiero agradecer a mi familia por todo el apoyo incondicional durante el doctorado y a lo largo de mi vida. Ustedes son mi motivación y ejemplo a seguir.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Chapter 1

Introduction

1.1 Perspectives on human phenotypes

A *phenotype* is defined as the set of measurable and observable traits in an organism [1], encompassing properties such as behavior, development and morphology. In humans, some immediate examples are height, weight, and eye color. Phenotypes exist not only at the macroscopic scale: RNA and proteins are also examples of phenotypes that are expressed at the molecular level.

The *genome* sequence of an organism is the collection of nucleotides (A, C, G and T in the case of DNA genomes) that make up the genetic material of a species, materialized in the form of *chromosomes*, for which humans have 23 pairs. Some regions of the chromosomes contain *genes*, which are regions of DNA that can encode molecular phenotypes. By means of mutations, different arrangements of these nucleotides can exist within a given gene, leading to variants called *alleles*. The combination of these alleles defines the genetic composition of an organism, also known as the *genotype*. *Environment* refers to the circumstances that surround an organism, such as weather, pollution, diet, and other biotic and non-biotic components. At all scales, the manifestation of a specific phenotype will be mediated by two main factors [2]:

$$\text{Phenotype} \leftarrow \text{Genotype} + \text{Environment} + \text{Genotype} \times \text{Environment}$$

1.1.1 Phenotypes and the quantification of biological units

A fraction of phenotypic variation is attributed to the genotype, especially with regards to an organism's morphology [3]. For many other traits, phenotypes and genotypes are not always correlated, and thus, they cannot be considered as equivalent. A component of the variation is also explained by environmental factors that play a direct role in how a phenotype is expressed. Finally, the interaction of both the genotype and the environment can lead to complex associations that regulate phenotypes: an example of a well documented interaction is how sunlight (environment) affects different skin tones (gene/genotype that regulates pigmentation) and their relationship with cancer risk (phenotype) [4].

Mapping the relations between genotypes and phenotypes (Fig. 1-1) is an important task since understanding which variants lead to which traits is paramount to understanding disease. For this reason, in recent years, a large amount of research has been devoted to completing this knowledge graph through Genome Wide Association Studies (GWAS), which seek to relate single nucleotide polymorphisms (SNPs) with disease [5]. By 2020, 186,000+ associations with almost 125,000 SNPs have been indexed in the NHGRI-EBI GWAS catalog [6].

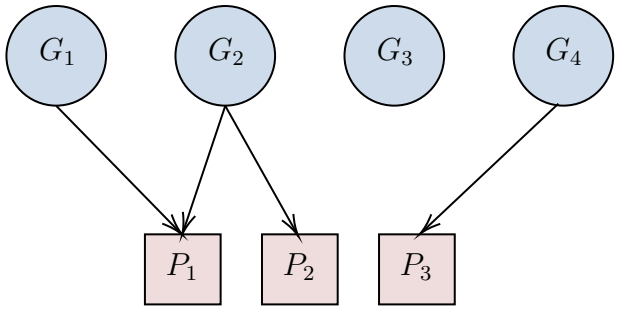


Figure 1-1: **Genotype-phenotype mapping.** A genotype G_i may influence a phenotype P_j, either individually or in conjunction with other genotypes.

One layer of abstraction above the genotype we find *gene expression*, which refers to the generation of a functional gene product (RNA or protein). To put this into context, let us remind that Francis Crick introduced in 1957 the central dogma of molecular biology [7], where the information flow between molecules in a cell is stated: from DNA to RNA and finally to proteins (Fig. 1-2). During the process of protein-

coding gene transcription, messenger RNA (mRNA) is produced. These molecules are the link between genes and proteins since they dictate how the latter will be synthesized. Thus, mRNA levels are used as a measure of gene expression.

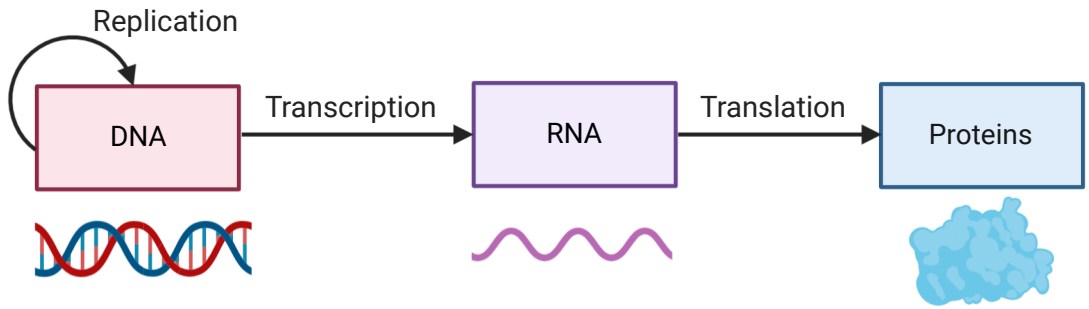


Figure 1-2: **Central dogma of molecular biology.** Crick's central dogma established that information in a cell flows from DNA (which is itself copied through the process of replication), then transcribed to RNA and finally translated into proteins.

In order to obtain a quantifiable representation of mRNA and other biological units, researchers rely on sequencing technologies, especially on Next-Generation Sequencing (NGS), which refers to a group of DNA sequencing technologies that superseded the Sanger DNA sequencing method that was prevalent three decades after its emergence in 1977 [8]. NGS techniques have changed the way in which biological research is performed since they allow to sequence a large amount of DNA fragments (in the order of millions) in a parallel way, which has led to speed increases to sequence full genomes at lower costs [9]. As an illustrative example, sequencing the complete human genome in 2019 was around 100,000x times cheaper than when compared to the cost in 2002 (Fig. C-1) [10].

Several applications are possible with NGS sequencing platforms: genome sequencing and resequencing, analyzing DNA-protein interactions, epigenome sequencing and transcriptomic profiling via RNA-seq [11]. Out of these techniques, RNA-Seq is of particular interest to this thesis work: with this approach it is possible to obtain precise transcript level measurements. These aid in the characterization of functional elements in the genome as well as in understanding the molecular composition of cells and tissues [12].

Turning raw sequencing data into an usable numerical form for statistical analyses

requires several steps. First, the *reads*, which are raw sequences that are generated when a sample is processed by a sequencing instrument [13] have to be aligned to the reference genome of the species of interest (note that it is also possible to assemble reads from species without a reference genome, which is a process called *de novo assembly* [14], and not covered here). Before actually performing this step, quality control statistics for the reads should be generated to assess if additional preprocessing is necessary before performing the alignment. Second, expression is quantified with summaries such as count tables that are generated at the level of the units of interest (genes, transcripts, exons, etc.) based on the mapped reads [15]. Due to the complexity of the process and to encourage analysis reproducibiltiy, bioinformatic pipelines such as GRAPE [16] aim to put together all the software tools required to go from raw data to quantification.

Gene expression quantifications usually require further normalization, since a part of measurement variability may be attributed to technical factors which should be accounted for in order to recover the true biological variation. Some of these factors are inherent to the sequencing process such as sequencing depth, mapping bias and errors. Other factors relate to the variability that might be introduced by different technicians performing the experiments. Lastly, gene length and sequence composition also need to be accounted for [17].

Gene length is particularly important to correct for, since this characteristic is directly linked to the estimation of gene abundances and has consequences when making comparisons across genes: a larger gene has a higher probability of capturing more reads during the sequencing process than a smaller gene. One of the first normalization methods that arose to account for gene length is RPKM, which refers to "reads per kilobase of exon model per million mapped reads" and was introduced by Mortazavi et al. [18]. Let r_{g_i} = reads mappings to the sequence of gene i, fl_g = sequence length for gene i and R = total of mapped reads for a library. RPKM for a gene g_i is then computed as:

$$\text{RPKM}(g_i) = \frac{r_{g_i} \times 10^3}{\text{fl}_{g_i}} \times \frac{10^6}{R} \tag{1.1}$$

where 10^3 normalizes for differences in gene length while 10^6 normalizes for differences in sample sequencing depth. Another unit related to RPKM is FPKM (Fragments per Kilobase of Transcript per Million fragments mapped), which is used for paired-end sequencing since a fragment is a pair of reads whereas initially reads were single-end [19]. RPKM and FPKM became widely used, but Wagner et al. [20] noticed that there were statistical biases present in RPKM normalization: average RPKM was similar among technical replicates but this was not the case across samples. They found that this was due to usage of the total number of reads in the normalization, which omitted the fact that $R/10^6$ is not a measure of the total transcript number but a proxy that will depend on the length distribution of the RNA transcripts across samples. For this reason, they proposed the transcripts per million (TPM) normalization:

$$\mathrm{TPM}(g_i) = \frac{r_{g_i} \times \mathrm{rl} \times 10^6}{\mathrm{fl}_{g_i} \times T} \tag{1.2}$$
$$T = \sum_{i=1}^{|G|} \frac{r_{g_i} \times \mathrm{rl}}{\mathrm{fl}_{g_i}}$$

where rl refers to the read length and G refers to the set of all genes measured in the experiment, with the rest of the values being the same as in the calculation for RPKM. The authors have shown that TPM is invariant among samples and is a proportional measure for average relative molar concentration. Since there is no consensus on a standard unit for transcript level quantification [21], all the three units are frequently reported in the literature [22], with a preference for TPM in the last years.

A different source of concern when analyzing gene expression quantifications relates to possibility of having *batch effects* in the data. This term is generally used to refer to sources of technical variation that are more on the side of experiment design and sample processing. Examples of batch effects include different processing times, data coming from different laboratories and generated with different sequencing instruments or by different technicians, among other factors which can be confounded with explanatory variables in the data [23]. Although the distribution of sample gene expression can be corrected with the normalization methods described above, they cannot fully account for batch effects [24].

Besides simple linear models, families of algorithms exist that attempt to deal with batch effect correction in one way or another. For example, Johnson et al. [24] introduced the *ComBat* methodology in the context of microarray data integration across studies, which relies on parametric and non-parametric empirical Bayes to perform batch correction in situations with small sample sizes. Other approaches such as surrogate variable analysis [25] focus on capturing components of expression heterogeneity that can distort true associations when examining gene expression.

1.1.2 Phenotypes and gene expression

Human disease can be the product of the interactions between genes and the environment [26]. To understand the complexity of diseases, their mechanisms of action, and consequently their causes, examining gene expression is vital, since sets of genes can behave differently in healthy vs. diseased individuals with respect to a particular pathology or phenotype: we call this *differential expression*. Highly expressed genes with tissue specificity can have associations to disease, as is the case of *MYH6, MYH7* and *TNNI3*, three genes which show expression specificity in heart and that are related to cardiac diseases [27]. These sets of differentially expressed genes (see, for example, Fig. 1-3) have the potential to be used as a signature for tasks such as diagnostic procedures, therapeutic target identification and phenotype characterization [28]. Although discrepancies in single genes can be linked to disease, co-expression networks can also provide valuable information which can be used, for example, to refine pairs of gene-disease associations from GWASs [29]. Gene expression data is not limited to the context of the disease: it can also be used to characterize transcriptomic dynamics during development and normal organism physiology (for example, to establish baseline profiles for normal tissues [12]).

In the past, microarrays were the most common method to generate measurements of gene expression levels [30] to perform differential expression and related analyses. Examples of early gene expression studies include the identification of cell cycle-regulated genes of *saccharomyces cerevisiae* [31], molecular profiling of breast tumors [32] and characterization of human lung carcinomas into distinct adenocarcinoma subclasses [33]. Microarrays have been replaced by RNA-Seq in the last decade, with the change propelled by the cost decreases discussed in Section 1.1.1.

However, genes are not expressed in the same way across cells. For example, although a human pancreatic cell and a keratinocyte have the same genetic content (DNA sequence), the difference in the cells' is at the expression level of their genes, which in turn may translate into functional and structural differences that are also mediated at some degree by epigenetics and other regulatory mechanisms that control

how genotypes are activated or inhibited. Gene expression is also dynamic: expression profiles can change over time with respect to the environment, development and other processes and therefore when gene expression is measured, it is a snapshot that is representative only of that specific time point and might or not generalize to other time points [34]. Since gene expression levels are directly measurable, they are also considered quantitative phenotypic traits themselves. Due to these reasons, a considerable amount of effort has gone into investigating cell-type specific gene expression and its relationship with other complex traits [35], including diseases. This has been possible due to the emergence of single-cell RNA sequencing (scRNA-seq), which is a sequencing method that allows to measure transcriptome-wide gene expression at the single-cell resolution, enabling the identification of cell-type clusters, hierarchies and state transitions [36]. Cell-type specific expression is discussed in Section 1.1.4.

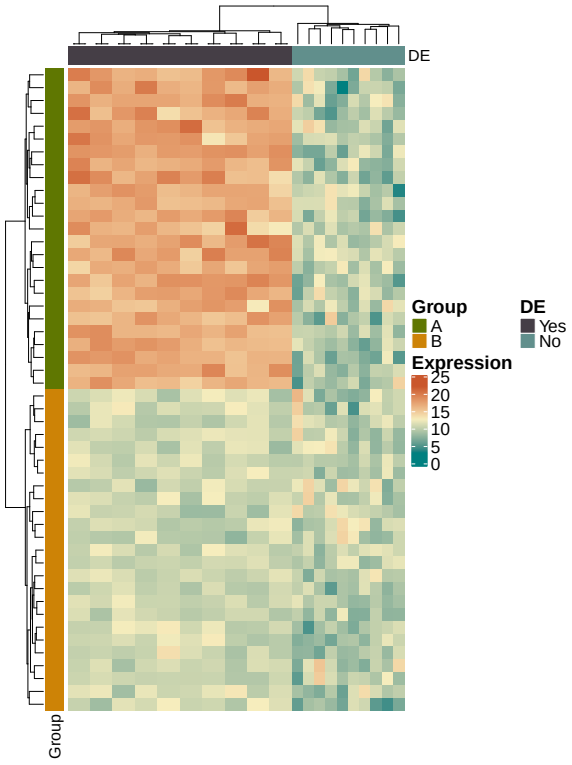


Figure 1-3: **Differential expression**. Illustrative example of a synthetic gene expression dataset where a subset of the genes (columns) show differential expression with respect to two groups of samples (rows).

The generation of these types of large datasets coupled with the increase of computational power has led the field of biology to also take part of the *big data* phenomenon that was previously prevalent mostly in physics and astronomy [37], with all the statistical benefits and practical disadvantages that come with processing these large volumes of data (see Section 1.1.3). Genomics datasets can also be useful for research purposes that are not directly related to the scope of the original research project that produced the data. This has led to the now commonplace practice of depositing high-throughput functional genomics data in databases such as the Gene Expression Omnibus (GEO) [38], ArrayExpress [39] and the Sequence Read Archive [40]. These are examples of repositories that function as archives that allow users to visualize and query relevant data for a particular topic of interest. The existence of such databases encourages data mining and knowledge discovery, meta-analyses by aggregating summary statistics across multiple experiments, and also fosters statistical analysis reproducibility. Nevertheless, these databases are not without caveats: data are usually generated via different protocols, metadata descriptions are usually not consistent across studies, reference annotations can vary across versions or even suffer from software-related encoding issues [41], among other problems that can make data integration a task that is not at all trivial [42].

1.1.3 Phenotypes and the curse of dimensionality

Transcriptomics datasets tend to be high-dimensional with a number of features (explanatory variables) p usually much larger than the sample size n. For example, GENCODE, which is an encyclopedia of genes and gene variants [43] reports in its latest annotation (version 35) a total of 60,656 genes in the human genome, out of which 19,954 are protein coding and 17,957 are long non-coding RNA genes. Reaching a sample size within a uniform, unbiased experimental setting that matches these numbers is still uncommon: as a point of comparison, The Cancer Genome Atlas (TCGA) has performed molecular characterization of approximately 10,000 human tissue specimens, but they are stratified over 33 different cancer types [44].

Identifying the effects of transcriptomic changes in the phenotypic space can prove difficult. In Section 1.1.1 we discussed how the genotype can affect the phenotype: to give an example of how large the genotype space is, consider that there are around 3×10^9 bases in the human genome, with each base being one out of four possible nucleotides. This means that there are at most $4^{3\times10^9}$ possible combinations of genotypes. Although this is a very simplistic calculation (since we would need to account for the valid moves in the genotypic space, like genetic rearrangements, sexual recombination, etc. [45]), even if we considered the fact that the existing variation between two individuals is around 0.6% of the total number of base pairs [46], we would still find that $4^{1.8\times10^8}$ is a very large number of possibilities.

To motivate why these large amounts of variables can be a problem in the context of predictive function estimation, let's consider the framework introduced in [47]: an input vector $X \in \mathbb{R}^p$ with an output variable $Y \in \mathbb{R}$ and a joint distribution $P(X, Y)$, with the goal of estimating a function $f(X)$ to predict Y. We first need to define a way to quantify the error of our predictions: this is done through a loss function $L(Y, f(X))$. A common example used in many statistical learning tasks is the squared error: $L(Y, f(X)) = (Y - f(X))^2$. The expected squared prediction error (EPE) for

f is then:

$$\text{EPE}(f) = E\left(Y - f(X)\right)^2 \tag{1.3}$$

$$= \int \left[y - f(x)\right]^2 P(dx, dy) \tag{1.4}$$

If we factor the joint density, we can rewrite this in terms of the conditional expectation:

$$\text{EPE}(f) = E_X E_{Y|X}\left(\left[Y - f(X)\right]^2 | X\right) \tag{1.5}$$

When this is minimized pointwise, we obtain the regression function. In other words, at a given realization x, the best prediction for Y (as measured by the mean squared error) will be given by the conditional mean:

$$f(x) = \text{argmin}_c E_{Y|X}\left([Y - c]^2 | X = x\right) \tag{1.6}$$

$$= E(Y|X = x) \tag{1.7}$$

Methods like k-nearest neighbors leverage this to make predictions: consider a point x, its k-neighborhood $N_k(x)$ (i.e. the k-th closest points) and μ being the mean; we can then approximate the expectation by simply averaging the y_i's over the data within a given region surrounding x:

$$\hat{f}(x) = \mu(y_i | x_i \in N_k(x)) \tag{1.8}$$

With a large sample size it is more likely that the $N_k(x)$ points are closer to any given point x, and the greater the k, the more stable the average. However, as p (data dimensionality) increases, the size of the neighborhood increases as well, making nearest neighbors quite unstable if the sample size is not very large, which, in practice, is usually the case. To intuitively understand this, let's consider a p-dimensional unit hypercube. In this hypercube, we can define a neighborhood that captures a fraction r of its volume with a given edge length e_p. This relation is given by $e_p(r) = r^{1/p}$. It is easy to observe (see Fig. 1-4a) that as we increase p, the hypercube edge

length required to capture the same fraction of volume r also increases: therefore, in higher dimensions, data is distibuted sparsely and local neighborhood estimation starts breaking down since we would need to cover a large part of the input range to be able to capture the same amout of volume than when compared to lower dimensions. Even if we reduced r we would observe large variances.

The effect of dimensionality on sparsity can already be observed in low-dimensional spaces [48]. To demonstrate this, we generate a set of 500 normally distibuted points (X, Y) with both coordinates sampled from $\mathcal{N}(\mu = 0, \sigma = 1)$ (see Fig. 1-4b). For the purpose of illustration, let's consider a neighborhood of 1σ around $(-1, -1)$. In two dimensions, only 18.2% of the points fall within 1σ of $(-1, -1)$, indicated with blue color inside the circle in the figure. When the data is projected independently on the x and y axes, 45.8% and 47.6% of all the points fall within 1σ of -1, a fraction that is considerably larger than when compared to two dimensions. With a higher number of dimensions, this percentage would rapidly decrease, as shown previously on Fig. 1-4a.

In practical terms, this means that whenever we want to estimate a phenotype (response variable) from any transcriptomic data type (for example, gene expression, or the genotype) for which the dimensionality is high, we would need an extremely large sample size to have an unbiased characterization of the joint distribution of the data. Assessing outliers becomes difficult due to sparsity, and other confounding effects like multicolinearity can occur. A related problematic is model complexity, especially *overfitting* (Fig. 1-5, third column), which can easily occur in the context of predictive function estimation: in short, this term refers to the model tightly fitting the noise (either technical or biological) that is part of each sample besides the underlying data relationships [49]. On the other hand *underfitting* (Fig. 1-5, first column) can also occur when the model is too simple that it fails to capture the generalities of the underlying true model. The relationship between underfitting/overfitting and model complexity will be formalized in Section 2.2.

However, not all is lost. Many methods have been developed to deal with the problem of many variables and few observations in statistical learning tasks: for example,

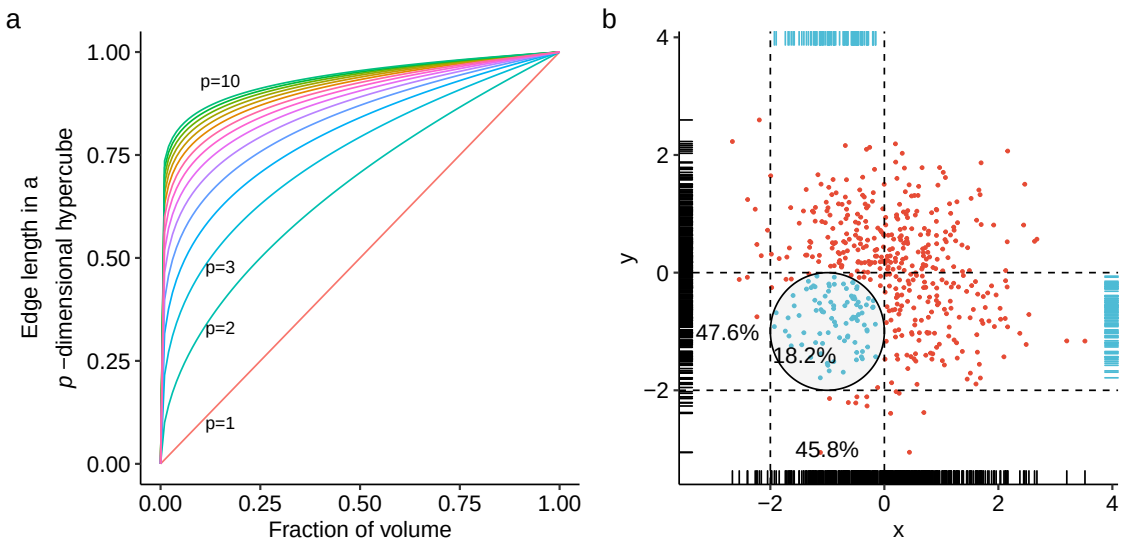


Figure 1-4: **The curse of dimensionality**. **(a)** Edge length of a p-dimensional unit hypercube (y-axis) needed to reach a certain fraction of volume (x-axis). Curves are shown for $p = 1$ up to $p = 10$. **(b)** 500 randomly-generated points, with each dimension (x and y) following $\mathcal{N}(\mu = 0, \sigma = 1)$. Marginal distributions are shown with black marks on the left and bottom sides of the plot. The subset of the marginal distribution for the points inside the circle is shown in blue marks on the right and top sides of the plot.

dimensionality reduction techniques, which are discussed in Section 2.1, encode the data into a reduced number of factors in such a way that the data properties are retained. Other common practices in the context of gene expression analysis (but also generalizable to other data types) include gene filtering while data preprocessing (for example, with minimum variance thresholding, removing genes that are lowly expressed, or also through more sophisticated wrapping and filtering methods [50]), gene weighting, gene ranking, and projection search. Thanks to these, it has been possible to perform discrimination of cancer types in leukemias, lymphomas and brain tumors [51] and in some cases yielding clear decision boundaries with a reduced number of genes.

Another different, but related type of problem, is the one of multiple comparisons and false positives. In the context of differential expression each gene is subjected to a statistical hypothesis test in order to determine if a difference exists in the expression of that gene between control and case groups. Each of these tests yields a p-value

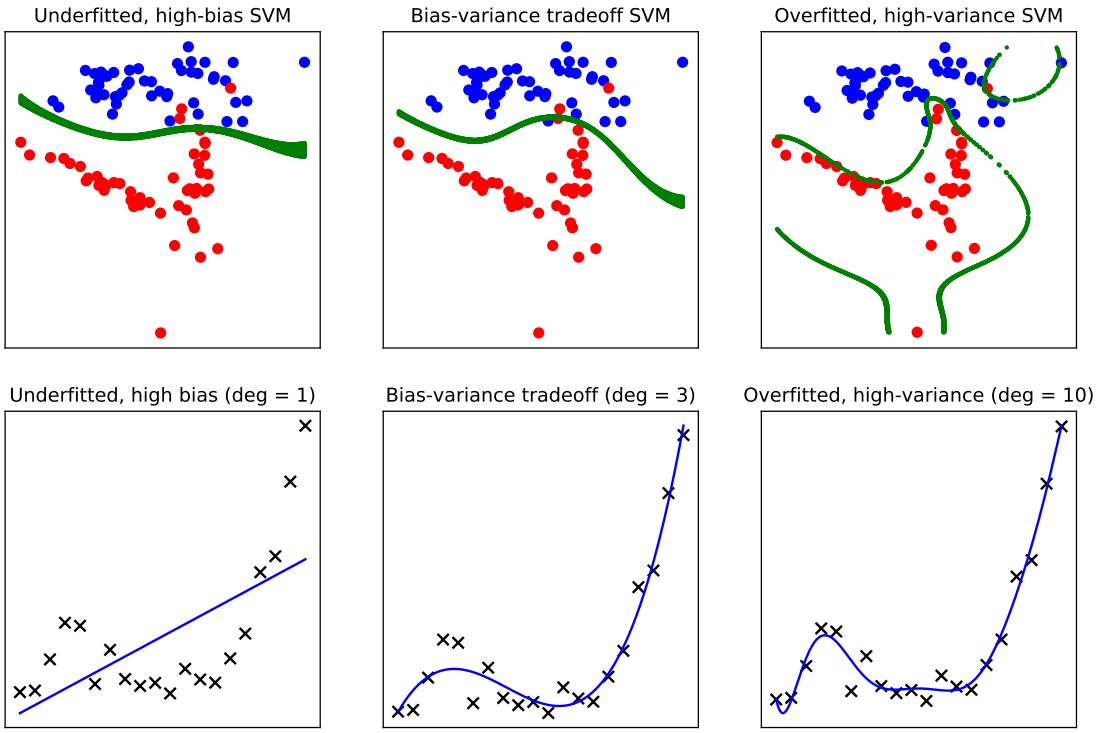


Figure 1-5: **Model complexity**. The top row shows a classification (2 classes) problem with a support vector machine (SVM) using a radial-basis function kernel; the decision boundary is shown in green. The trade-off between the bias and the variance is controlled through the C parameter in the SVM model (increasing, from left to right). The bottom row shows a regression problem (polynomial function estimation). The trade-off between the bias and the variance is controlled through the degree of the adjusted polynomial (increasing, from left to right).

which is associated with the specificity of that test (e.g. with the prevalent significance level of $\alpha = 0.05$, the specificity of the test is 0.95). Many genes are tested at the same time, and with a lenient significance level, false positives are bound to happen. This can be illustrated very easily: if we want to test as few as 50 genes for differential expression at $\alpha = 0.05$, the probability of observing at least a significant result by chance is already quite large:

$$\begin{aligned} P(\text{at least one significant}) &= 1 - P(\text{none significant}) \\ &= 1 - (1 - 0.05)^{50} \\ &\approx 0.923 \end{aligned} \tag{1.9}$$

This is unacceptable, considering that the objective of these tests could possibly be to provide information towards diagnostic testing. To this end, multiple adjustment procedures have been developed [52]: one of the first is the Bonferroni correction, which simply adjusts the significance cutoff to be α/n where n is the number of tests. In the example above, the Bonferroni-adjusted significance cutoff would then be 0.001, and the probability of at least one significant would then be close to the original level of α: $1-(1-0.001)^{50} \approx 0.0488$. However, this adjustment method is considered too strict [53] and other methods are more widely used in bioinformatics instead, such as the False Discovery Rate (FDR) which aims to control the expected proportion of incorrect rejections of the null hypothesis [54].

In summary, classical statistics and statistical learning methods are powerful tools to extract biological knowledge from large databases. Both have withstood the test of time, however, domain expertise will always be necessary in the interpretation of the results: even if a specific hyphotesis or model is numerically sound, it does not necessarily mean that it is biologically plausible. The issue of high-dimensionality presented here is inherent to all the case studies that compose this thesis work, although treated in a different way in each chapter.

1.1.4 Inferring cellular composition from gene expression

When analyzing bulk RNA-seq data from tissues, one must consider the fact that the measured gene expression levels represent a tissue average, since many different cell types can compose a tissue, and genes can be expressed differently in each of these cell types. Thus, a possibility exists that when we observe differential expression in a gene, the change in expression is confounded only because the underlying proportions of cell types are not uniform across samples, for example, with the abundant cell types masking the effect of those that are not as abundant [55]. This issue can be particularly troublesome in the case of cancer, where tumor cell type composition is quite often not homogeneous and can affect treatment outcomes [56] and can be quite different when compared to healthy tissues. In an attempt to correct for cell type abundances when performing gene expression analyses, as well as to link cell type abundance estimates with phenotypic traits, a series of methods have been developed to computationally estimate the levels of these cells from bulk RNA-seq data. This problem can be conceptualized and approached from two different angles: deconvolution and enrichment.

In cell type deconvolution, we assume that gene expression for a sample is a linear combination of expression values of each cell type present in the sample. In other words, this is:

$$G = SP \tag{1.10}$$

where $G \in \mathbb{R}^{p \times n}$ is a gene expression matrix for p genes across n samples, $S \in \mathbb{R}^{p \times k}$ is a cell-type specific "signature matrix" that contains average expression values for each gene-cell type pair, and $P \in \mathbb{R}^{k \times n}$ is the matrix of mixing proportions for k cell types across the n samples. There are two main approaches to deconvolute G: supervised and unsupervised [57]. In the supervised approach, G is available, but also either S or P and thus we are concerned with the estimation of only one matrix. Linear and ordinary least squares methods are commonly used to minimize the differences (sum of squares) between G and SP:

$$\min_{S \text{ or } P} \|G - SP\|^2 \tag{1.11}$$

Note that this is stated as an unconstrained optimization problem, and as such, a proportion for a cell type i in sample j can have $\text{sign}(p_{ij}) \in \{-1, 0, 1\}$ which is not entirely sound since a negative proportion does not have a biological meaning. For this reason, a non-negativity constraint and a sum-to-one constraint are enforced while estimating the vector of proportions for each sample. This is a convex quadratic programming problem and can be solved, for example, via non-negative least squares (see, for example, [58]) or constrained least squares [59]. Here, we proceed to illustrate an example of the latter with a synthetic gene expression dataset.

Consider a matrix $G \in \mathbb{R}^{100\times150}$ of bulk RNA-seq gene expression levels (Fig. 1-6a) with the expression for each sample being an underlying mixture of 6 cell types. In order to generate G, we need to specifiy a true model for $S \in \mathbb{R}^{100\times6}$. The gene expression for each cell type in S is generated as $\mathcal{N}(\mu, \sigma = 0.5)$, with the cell type means specified in Fig. 1-6b. Finally, we generate the matrix of proportions $P \in \mathbb{R}^{6\times150}$ partitioning the samples in three groups, with each group showing specificity exclusively for two cell types, while the rest of the cell types have a a proportion of zero (Fig. 1-6c). Within each group of samples, the proportions for the first cell type (i.e. cell types 1, 3 and 6) were generated as $\mathcal{U}(0.5, 1)$, while the second cell type was generated as the complement so that the proportions for both group-specific cell types added up to 1. Finally, G is constructed as the matrix product of the cell type expression signature and the mixing proportions, plus an error term:

$$G = SP + E \text{ with } E_{\cdot j} \sim N(\mu = 0.1, \sigma = 0.5) \tag{1.12}$$

Most constrained least squares implementations (for example, `lsqlincon` from the `pracma` R package [60] which is used in this example to perform deconvolution) are not vectorized, and thus, here we reformulate the problem at the sample-level for the sake of consistency. Let $\mathbf{g}_i = [g_1, \ldots, g_{100}]^T$ be a column vector representing a sample in G and $\mathbf{p}_i = [p_1, \ldots, p_6]^T$ a column vector in P with the mixing proportions for that sample.

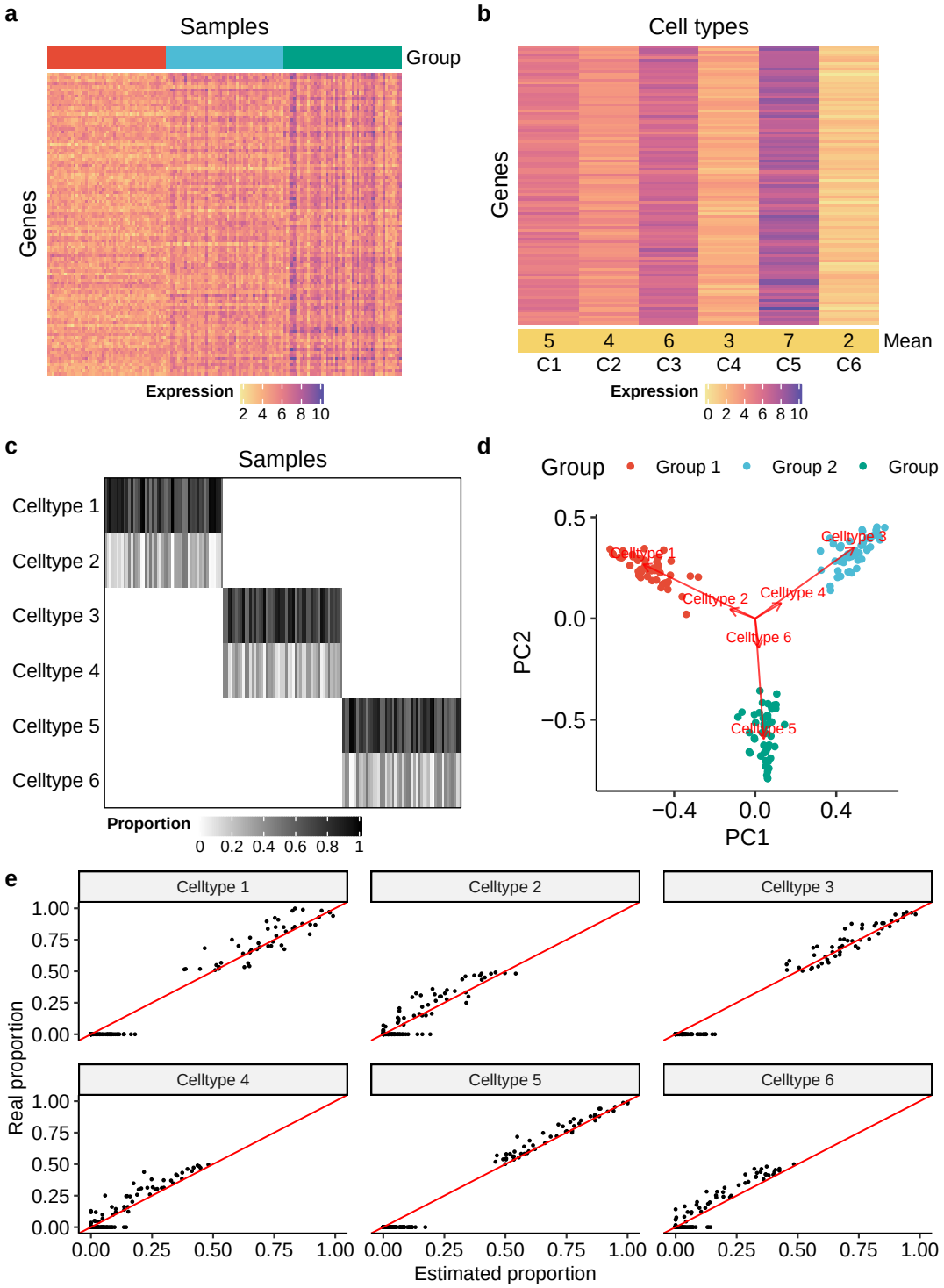


Figure 1-6: **Cell type deconvolution**. **(a)** Gene expression matrix, with samples divided into three groups, with each group having a mixture of two cell types. **(b)** Signature matrix for the cell types. **(c)** True cell type proportions for 6 cell types in 150 samples. **(d)** PCA over the deconvoluted cell type proportions for all the samples. **(e)** Real vs. estimated proportions for each cell type across all samples.

With the problem set up, we now proceed to perform deconvolution for each sample $i \in \{1, \ldots, 150\}$ by setting constraints over (1.11):

$$\begin{aligned} &\min_{P} \|\mathbf{g}_i - S\mathbf{p}_i\|^2 \\ &\text{subject to } A\mathbf{p}_i = 1 \\ &0 \leq \mathbf{p}_i \leq 1 \end{aligned} \tag{1.13}$$

where $A = \mathbf{1}$ is used to enforce linear equality constraints over the proportions, while the last constraint enforces their non-negativity. Afterwards, we can verify that cell type specificity is reflected adequately on the estimated proportions across the samples by performing PCA over $\hat{P}^T$ as shown in Fig. 1-6d, where it can be seen that the three groups of samples are separated, as expected, with the corresponding cell type loadings associated to each group. Note that within each group, the angle between the cell type loadings is close to zero since in this toy example one cell type is dominating and the other one is linearly dependent as they must add up to 1. On Fig. 1-6e the estimated vs. real (simulated) proportions are shown for each cell type.

Other supervised methods, such as CIBERSORT [61] which is based on support vector regression, minimize an objective function that besides computing a loss function also penalizes model complexity through the incorporation of a regularization term. In the unsupervised approach, also called *reference free* deconvolution, the estimation of S and P is performed at the same time using only G as an input. Unsupervised methods include non-negative matrix factorization (NNMF) which tends to be quite sensitive to the initial seed values for S and P [62], and Bayesian frameworks that tackle the problem through different formulations (see [63] for an example). With unsupervised approaches, it is necessary to perform additional analyses to map each deconvoluted component to an existing cell type by analyzing the marker genes that drive each component. In general, supervised methods tend to have a better performance than their unsupervised counterparts [57].

In the framework of enrichment, as opposed to deconvolution, we do not consider the cell type abundances as a composition that needs to add up to 1, therefore, any

estimated value is not considered a proportion but rather just a value that can be used to compare cell type abundances between samples. An example of such a method is xCell [64] which aggregates sets of marker genes (i.e. genes that are characterisically expressed in a given cell type) for 64 different celltypes using six different data sources. More than one set of marker genes can be associated with any given cell type. Of note, only the gene name identifiers are used without relying on any actual expression value. In Fig. C-2, a hierarchically-clustered heatmap is shown with the overlap of the marker gene sets (the marker list for each cell type is considered here to be the union of all marker sublists for that cell type) for each pair of cell types A and B as measured by the Jaccard index $J(A, B) = |A \cap B|/|A \cup B|$, where it can be seen that related cell types tend to share marker genes. In xCell, an enrichment score for each cell type marker sublist within each cell type is calculated using single-sample Gene Set Enrichment Analysis (ssGSEA) [65], which, in short, is a score calculated based on the position of the cell type markers within the sorted gene expression vector for a given sample. In other words, if the cell type is abundant within the sample, the marker genes are expected to be near the top of sorted gene expression vector. Finally, a raw enrichment cell type score for a sample is calculated as the average ssGSEA score across the cell type marker gene sets for a given cell type.

Although deconvolution and enrichment methods are a reasonable way to control for cell-type specific expression in bulk RNA-seq, gold standards of the cell type proportions are rarely available, and thus, the estimated cell type levels need to be validated with orthogonal measures. In relation to this, generating experimental data for single cells began in 2009, when Tang et al. [66] made the first description of a single-cell transcriptome analysis characterizing mouse cells in early developmental stages. Since then, flow-activated cell sorting (FACS) has become the most common strategy to isolate purified single cells, but other strategies such as laser capture microdissection and microfluidic technologies are also used [67]. Single-cell RNA-seq has been used to understand the role of gene expression in cellular states and functions both in healthy and diseased organisms, as well as delineating cell lineage relationships [68]. Identifying outlier cells has also been pointed out as means to aid in

characterizing drug resistance and relapse in cancer [69]. In terms of computational experiments, cell-type specific gene expression signatures derived from scRNA-seq have also been used to perform deconvolution in bulk RNA-seq data [70].

With the goal of briefly illustrating the initial steps that can be performed as part of an exploratory single-cell RNA-seq gene expression analysis to determine cell-type organization, here we use data generated by Braga et al. [71] that was used in their work to characterize the cellular landscape of the respiratory airways and lung parenchyma in healthy human lungs. In this example, we analyze a subset of 10,358 cells. First, we require a minimum number of expressed genes per cell (here, 200 genes) and filter out those cells that have a more than a certain threshold of their reads (here, 20%) coming from the mitochondrial genome (i.e. mitochondrial genes); this is done since low quality or dying cells usually present mitochondrial contamination. Second, the expression of each cell is normalized by the total expression, multiplied by a scale factor and then log-transformed. Third, with the aim of keeping the most informative genes and reducing the dimensionality, the top 2000 most variable genes across the cells are kept. Fourth, the expression of each gene is scaled so it has a mean of 0 and variance of 1 across cells, to avoid highly-expressed genes from dominating. Fifth, t-SNE dimensionality reduction (see Section 2.1.2) is performed over the processed data matrix, as shown in Fig. 1-7a-b. In those panels, cells are colored by the airway region of origin, and by the cell type labels devised by Braga et al., respectively. Note that in an exploratory analysis, the ground truth label for the cells is usually not known, and different techniques (for example, examining PCA loadings) are used to identify marker genes that might be related with a specific cell type, similar to the case of enrichment methods that keep a list of cell type marker genes. Finally, since the expression of each gene is known at the cell level, it is possible to visualize if specific cell type clusters show or not specificity for a given gene (Fig. 1-7c-d). Pipelines such as Seurat [72], which was used in this example, integrate the common operations that are involved in a single-cell analysis, such as feature scaling and normalization, metadata storage, dimensionality reduction and clustering, visualization, as well as more complex processes to combine multiple datasets among other tasks.

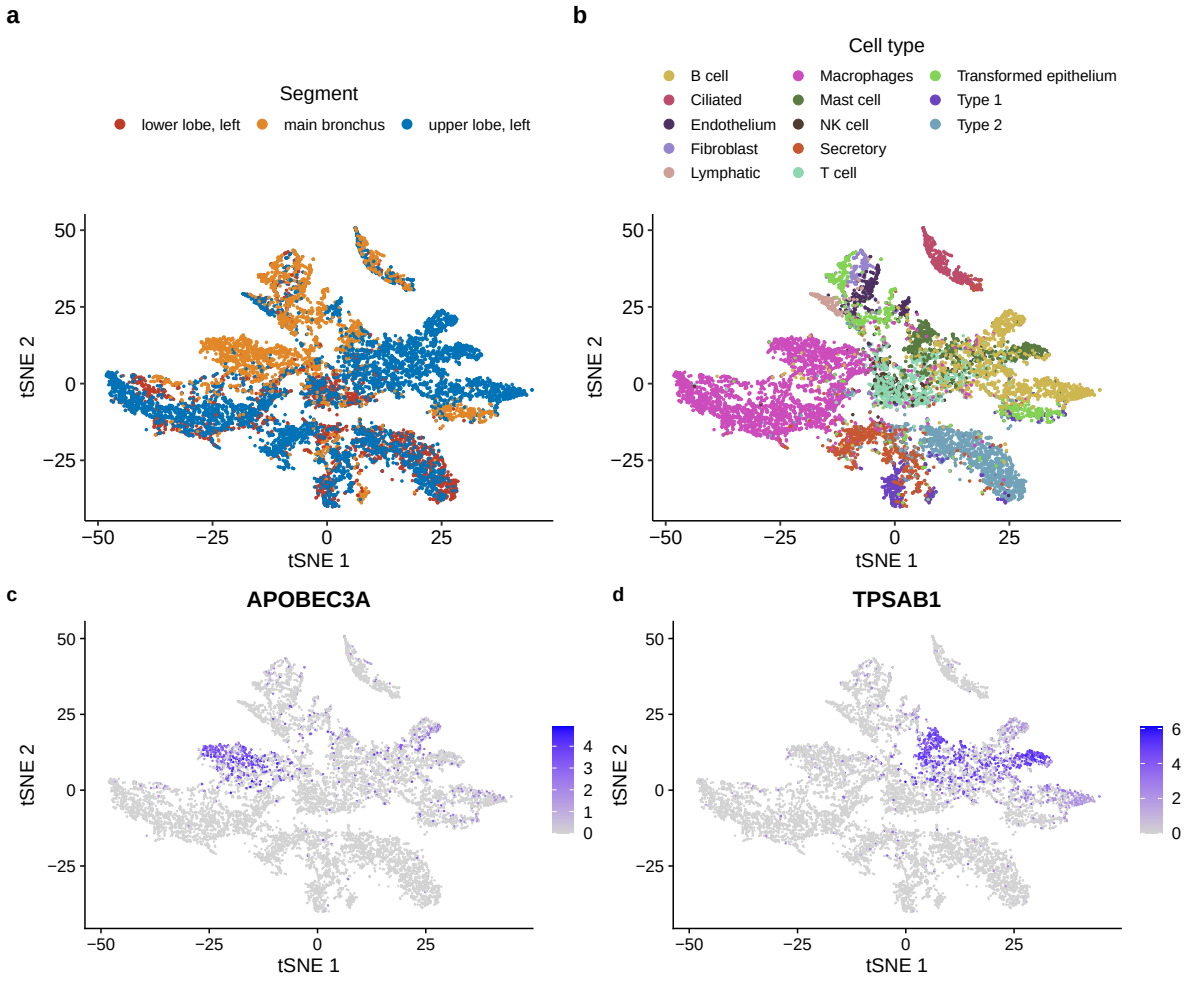


Figure 1-7: **Single-cell RNA sequencing profiling of lung cells**. In all the panels, t-SNE based on single-cell expression of 16,147 genes across 10,358 cells is shown. **(a)** Cells colored by the respiratory airway section of origin. **(b)** Cells colored by their inferred cell type clusters. **(c)** Expression of *APOBEC3A* at the cell level. **(d)** Expression of *TPSAB1* at the cell level. Figures produced with single-cell gene expression data generated by Braga et al. [71].

In Chapter 4 we address the problem of estimating cell type enrichments and associating them to histological phenotypes that correlate with disease states, as well as analyzing single-cell data from *Mus musculus* to demonstrate similarities with humans with respect to cellular composition clusters. Sex differences in human cell type abundances are covered in Chapter 5.

1.1.5 Quantifying biological knowledge from text

Biological data that can be mined for research is not limited to that generated by massive parallel sequencing technologies. Besides the large amounts of medical literature, patients' clinical data is often available as unstructured text reports [73] that are commonly stored in electronic health records. These can be leveraged, for example, to study disease co-morbidities and to track clinical outcomes on large cohorts of patients.

An important hurdle to overcome when mining clinical data is that the information is usually not readily available in a tabular form or organized in any other way that facilitates statistical analyses. Instead, it tends to be stored as free-form text that needs to be parsed in order to identify and annotate concepts of interest such as medical phenotypes, clinical symptoms, pathology review comments from tissue samples, radiological findings, among many others. Transforming these into a structured database is not trivial, since there are many issues that can arise: typographical and other random text errors, sparseness, high dimensionality, text incompleteness, biases [74], as well as interoperability issues in dataset integration [75].

Knowledge extraction and annotation of health records has been conventionally performed through manual review by clinicians, but recently, machine learning techniques such as Natural Language Processing (NLP) have been successfully used to develop pipelines with the goal of providing support for decision making in clinical practice, for example, in the estimation of coronary artery disease risk [76] and detection of adverse drug reactions [77].

The task at hand will determine how to extract and/or encode text into numerical features, as well how to use these for further analysis. For example, simple statistical features from a text corpus such as n-grams (which are a contiguous sequence of tokens) can be used to identify keywords in a text with the goal of document classification [78]. Other approaches rely on the creation of numerical vectors to represent text; bag-of-words is a one of the first methods that arose to do this, and consists in summarizing a sentence with a vector that counts the occurrences of each token in the

sentence, with the word position given by a dictionary of possible words occuring in a corpus of text. Examples of n-grams and bag-of-words are shown in Fig. 1-8. The bag-of-words approach, however, has the issue of generating very high-dimensional and sparse representations and for this reason, newer methods (as in [79]) make use of dimensionality reduction or neural networks to generate mappings that encode text into compressed lower-dimensional embeddings.

```
Sentence 1: Adenocarcinoma is a type of cancerous tumor.
Sentence 2: Glioblastoma is a type of brain cancer.

1-grams of Sentence 1: Adenocarcinoma, is, a, type, of, cancerous,
tumor
2-grams of Sentence 1: Adenocarcinoma is, is a, a type, type of,
of cancerous, cancerous tumor.

Bag order: ['Adenocarcinoma', 'Glioblastoma', 'a', 'brain', '
cancer', 'cancerous', 'is', 'of', 'tumor', 'type'
Bag 1: [1, 0, 1, 0, 0, 1, 1, 1, 1, 1]
Bag 2: [0, 1, 1, 1, 1, 0, 1, 1, 0, 1]
```

Figure 1-8: **Text feature encoding.** Examples of n-grams ($n = \{1, 2\}$) and bag-of-words. In the bag-of-words example, the length of the vector is determined by the union of the sets of tokens in the corpus (here, a set refers to the unique tokens of each example sentence). Then, each sentence is a transformed into a vector of counts for each of the possible words in the dictionary.

Other language processing tasks are more concerned with the identification of the parts of speech in a text corpus taking into account linguistic knowledge about the definition of the token and its context, as well as the syntactic dependencies between these tokens. An example of this is shown in Fig. 1-9, where the tagging is performed through a multi-task convolutional neural network pretrained on a large text corpus that encompasses different genres of text, including structural information and semantics [80].

In this line, there are specific NLP systems for knowledge extraction tailored to a particular domain instead of using a generic model like the one mentioned above. An example of this is cTAKES (clinical Text Analysis and Knowledge Extraction System)

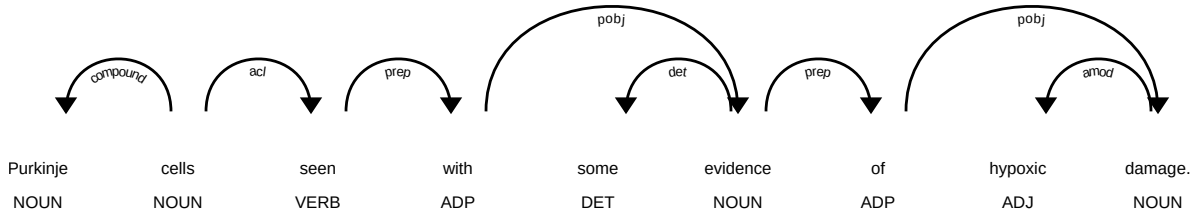


Figure 1-9: **Part-of-speech tagging.** Example of part-of-speech tagging for each token in sentence describing histological findings on a brain sample. The syntactic dependencies (arrows) between each token are also shown.

[81] that addresses text parsing tailored to the clinical domain. Besides part-of-speech tagging, these systems can perform tasks including, but not limited to, named entity recognition, rule-based tokenization, phrasal chunking, dependency parsing, negation detection, among many others.

In biomedical sciences (and also in other domains), issues can arise when trying to automatically interpret concepts, since these can be encoded in different ways depending on the knowledge subdomain, as well as due to variability attributed to whoever is tasked with encoding these concepts. To aid in the interpretation and translation of concepts between different terminology systems, compendiums of controlled vocabulary have been created. An example of this is UMLS (Unified Medical Language System) which is an ontology that aggregates biomedical vocabulary which also serves as a thesaurus [82]. Ontologies can serve as a base for uniformizing text and ensuring clinical texts are properly encoded: correct representations have been shown to lead to increases in machine learning classifier accuracies [83].

In Chapter 4, histological phenotype retrieval from pathology texts in free-form is performed, together with their usage in downstream analysis to associate changes in cellular composition with disease.

1.1.6 High-dimensional phenotypes: histological images

The field of histology is concerned with studying tissues and their microanatomic composition, focusing on cell structure and arrangement and their relation to organ function. To this end, tissue sections have to be prepared in such a way that its structural features are preserved for their posterior visual examination through the use of microscopes. The preparation usually involves i) fixating the tissue in a solution that allows the preservation of cell structure by inactivating degradative enzymes, ii) removing water from the tissue using alcohol solutions, iii) removing the dehydration solutions used in the previous step, iv) placing the tissue in paraffin within a mold to allow it to harden, and v) trimming/sectioning the paraffin block using a microtome to expose the tissue. Cells and their extracellular matrix usually do not have color, and thus, a dye has to be applied to distinguish tissue components. This process is called staining, and one of the most popular methods is hematoxylin and eosin (H&E) which stains cytoplasmic structures and collagen in pink, while DNA in the cell and other elements are colored with a purple/dark blue tint [84]. Tissue characterization through histopathology is an important diagnostic tool in detecting and diagnosing cancer and its malignancy level [85] as well as to identify infectious organisms [86].

The first attempts that relied on numerical methods to quantify cell morphology heavily depended on human intervetion and parameter tweaking to identify pattern classes (for example, to distinguish cell types). An example is shown in Fig. 1-10, where Prewitt et al. [87] used a cytophotometer to record scanned fields as matrices of grayscale optical density values. By analyzing the frequency of the optical density in these matrices (Fig. 1-10b-c), it was possible to determine boundaries for the background, cytoplasm and the nucleus. In this work, the idea of concatenating sets of parameters (contrast, nuclear area, cytoplasmic area) as numerical vectors was also explored in order to perform cell type discrimination in a multidimensional space.

With the increase in computational power and decreasing costs of data storage, pathology departments around the world are switching towards digital imaging techniques (as opposed to static image acquisition of a specific region) for image storage

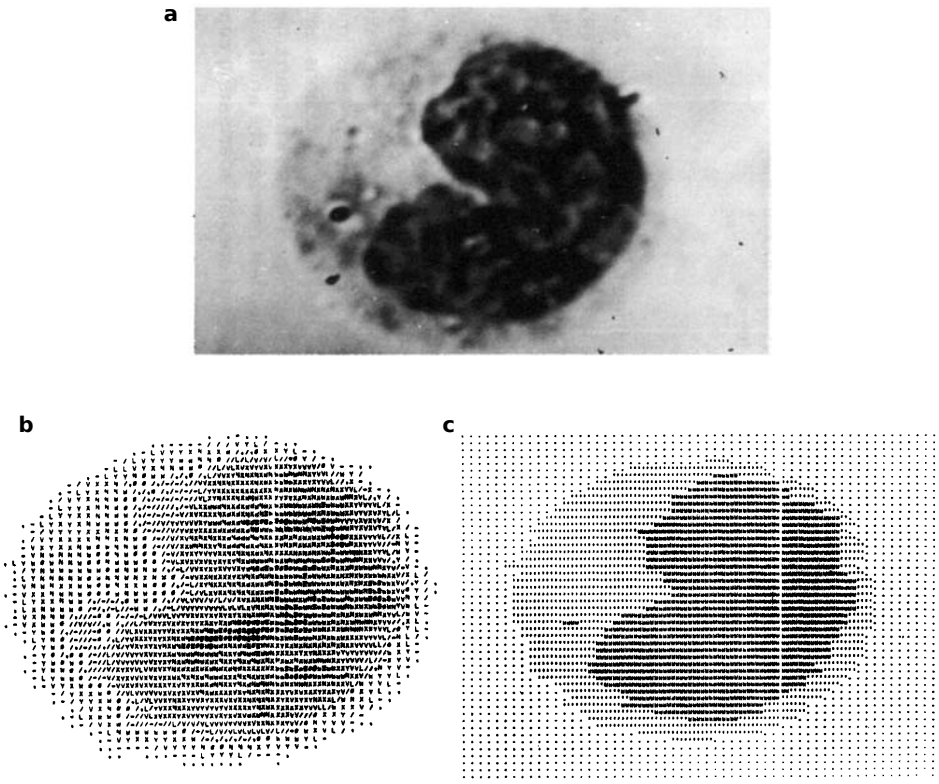


Figure 1-10: **Early analysis of cell images (a)** Photomicrograph of a monocyte. **(b)** Computer printout of (a) using a 32-level grayscale scheme. **(c)** Local minima thresholding of (b) to generate a 3-level grayscale image. Image adapted from Prewitt et al. [87].

and analysis [88]. The term *whole slide imaging* is used to refer to the digitization of glass slides of tissue sections so that they can be explored computationally, as if emulating conventional light microscopy. Digitization of slides through whole slide scanners started around the 1990s [89]. It has been pointed out to be superior in several respects when compared to their printed/static counterparts, especially in terms of image resolution and their annotation and data mining capabilities [90], as well as facilitating data sharing among different physical locations.

Image files generated by digital whole slide image (WSI) scanners are of high-resolution (determined mainly by the microscope's objective), typically resulting in files that can be up to several gigabytes in size [88]. Unfortunately, the digital pathology community has not yet adopted a standard format to encode these WSIs, even if open-source formats have already been proposed [91]. Proprietary file formats

designed by scanner manufactures are still the norm, with each format supporting different metadata fields, making interoperability difficult and expensive to support from the application perspective. As a consequence, software to analyze and process WSIs also tends to be vendor-specific. Open source libraries such as OpenSlide [92] aim to mitigate this issue by integrating support for several proprietary formats, making the low-level operations needed to read and process WSIs transparent to the end-user. Examples of popular WSI formats include Aperio (.tiff, .svs), Hamamatsu (.vms, .ndpi) and Leica (.scn). In Section 1.2.3 an example of the pyramidal image structure used by the Aperio SVS format is described.

Image analysis techniques have advanced considerably since the example shown in Fig. 1-10. Many of the first advances were centered in feature extraction methods based on human intuition. The way we, as humans, interpret images is object-based in the sense that we associate specific visual patterns with specific concepts. As such, we can answer questions such as: how many cells are in the image, what is their size and distribution? In computer vision, the representation of an image is pixel-based and there is no inherent meaning to the pixels that compose it. But statistical descriptors of the morphological appearance (of the human conceptualization) of these objects can be derived, for example, cell areas, bounding boxes, perimeter, radii, hue, distances, etc. Besides the object descriptors, structural information about cellular arrangements can also be quantified with graph-theoretical metrics such as the degree of a node or betweenness centrality (where the node is the object of interest, for example, a cell) that convey information about how other nearby cells are arranged with respect to a specific node. Sets of features related to spatial arrangements can be generated through structures such as minimum spanning trees, k-NN graphs, among others [93]. These features are termed *handcrafted features* since the selection of the objects to characterize and the process to generate summary statistics for these objects is human-guided [94]. The extraction of these features (Fig. 1-11a) was considered to be an independent problem to that of learning a model that seeks to solve a specific task (Fig. 1-11b).

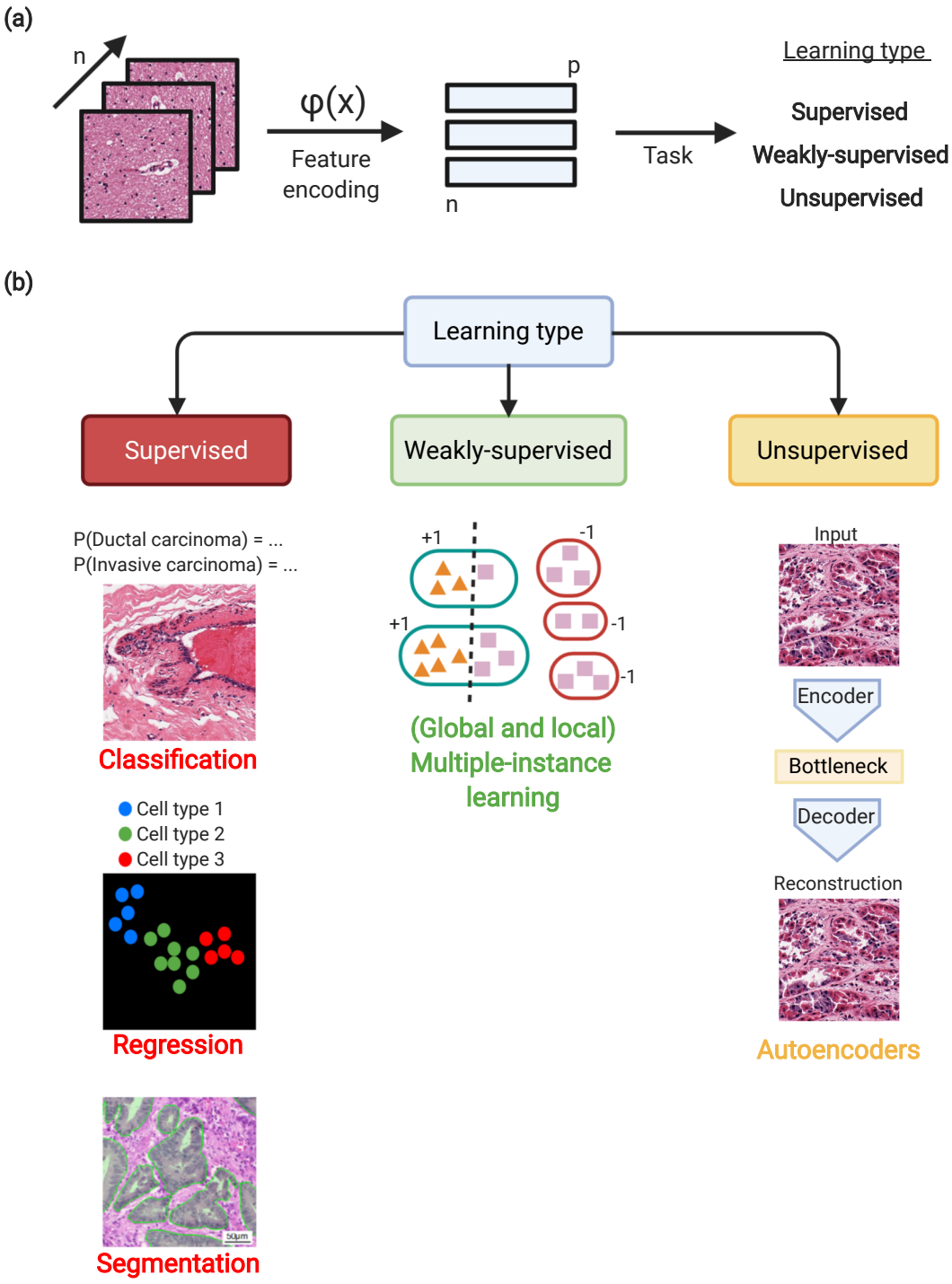


Figure 1-11: **ML-based histopathology image analysis.** **(a)** A region x of a WSI is encoded into a numerical representation $\mathbb{R}^p$ by a function $\phi(x)$. These feature vectors can then used for specific tasks. **(b)** Overview of common learning types in histopathology image analysis. Supervised methods include classification, regression and segmentation; weakly-supervised methods are mostly based on multiple instance learning, and unsupervised methods are based on the general principle of autoencoders. The segmentation image was retrieved from [95], while the rest of the images are of samples from the GTEx project [96].

Although handcrafted features can still be useful for several applications, feature extraction and model development have since mostly merged into a single step through the usage of deep (convolutional) neural networks. When training a deep learning model to optimize a set of parameters, input images are implicitly converted into lower-dimensional representations that are generated in such a way that a specific loss function is minimized, obviating the need of human intervention to decide which image patterns are of relevance to solve a specific task. The quantification of histological images in the form of feature vectors, either handcrafted or through automated means, can then be used in many different tasks (see Fig. 1-11b). At the applied level, these include, but are not limited to, cell detection, cancer classification and grading, tumor segmentation and survival prognosis. An expected benefit of these computational models is to reduce the inter-pathologist variability that might occur when evaluating an histological image, for example, when using the Gleason score or the Nottingham score to assess prostate and breast cancer, respectively [89].

As discussed before, WSIs at their native resolution have very large dimensions which are usually in the order of 10,000 to 50,000 pixels for either coordinate (width or height). Such large sizes make the processing of a WSI as a single unit intractable: current deep learning architectures already tend to be computationally costly in terms of parameter storage at training time when processing images that are much smaller than WSIs, rendering the direct usage of the latter infeasible. For this reason, many of the current deep learning methods and their applications rely on the extraction of smaller-sized image tiles (also called patches) from the WSI [97]. Depending on the application, a random sampling of tiles from the WSI might be sufficient, otherwise, the slide can be divided into a grid of tiles and all the tiles corresponding to tissue foreground can be kept. In Chapter 6, we further describe the problem of tile extraction from WSIs, and a software solution is presented to automatically extract tiles from the foreground (tissue slices) of the WSI.

The literature covering other approaches that do not completely depend (at least towards the end-stages of the learning task) on image tiles is still scarce. Most of these alternative approaches focus in exploting the spatial relationships between

the tiles (see, for example, [98]) or performing meta-learning with tiles extracted at different resolutions of the WSI. Attention models that rely on dynamic selection of tiles without the need to load the complete WSI into memory have also been explored [99].

Another relevant aspect to consider is the availability of image annotations, at least for the applications that require it, which mostly fall into the realm of supervised learning. Oftentimes, these ground truth annotations are only available at the WSI level. For example, a WSI can be marked as "healthy tissue" or "diseased tissue", but no further information is available about which particular regions of the WSI are affected. On the other hand, patch level annotations (for example, creating masks for specific cells, or delineating the boundary regions of a tumor) are most commonly used in strongly supervised settings, but these types of annotations are more scarce since they are costly: pathologists have to manually annotate the regions of interest in each patch.

Now that the key considerations have been laid out, we can proceed to describe the fundamental aspects of each of the learning tasks in (Fig. 1-11b), following the scheme presented in [89]:

1. *Supervised learning*: in this category, each observation is associated with a ground truth label that can exist at different levels (WSI, tile, or pixels), and the goal is to predict the label. Three main paradigms exist for using labeled data. Although an applied task can be (and often is) shared across these, the way that the problem is formulated will vary, in the sense of what the network architecture and objective function are and how the latter will be optimized:

    - Classification: At the global level (WSI or tile), it can be used to decide if the unit of interest is affected by a specific pathology, for example, cancer. Alternatively, at the local (pixel-based) level, common tasks include cell and mitosis detection through sliding-window approaches. Most commonly addressed using convolutional neural networks (CNNs) and derivative architectures.

- Regression: commonly used to perform object detection and localization (predicting a position) of specific objects in an image. Can be used, for example, for cell and mitosis detection. As opposed to classification, in the regression approach pixel constraints are enforced to take into account the neighborhood with respect to a specific pixel (which is commonly the center of the object of interest). Regression tasks can also be about predicting the severity score of a specific pathology. Most commonly performed at the tile level, and modeled using CNNs, fully-connected networks (FCNs) and derivative architectures.

- Segmentation: in semantic segmentation, the goal is to generate biologically meaningful masks to delineate histological primitives. Commonly addressed at the tile level through FCNs and U-Nets [100].

2. *Weakly-supervised learning*: refers to the inference of granular annotations (at the tile or pixel level) from the annotations/labels available exclusively at a less granular level, which is usually the WSI. For example, a label of either "diseased" or "healthy" can be assigned to the WSI, but it is not known exactly which regions contribute to the "diseased" label. With this paradigm, annotations for granular regions that contribute to the WSI label can be inferred. This can potentially be of help to pathologists who need to perform manual annotations. The standard methodology to perform this is *multiple-instance learning* (MIL) [101], and although several specific subdefinitions exist, the general idea is based on the concept of a "bag", which refers to a set of "instances" (either tiles or pixels) that are extracted from a WSI that has a label available only at the coarse level. The instances will inherit the label of the WSI they are extracted from. For example, in the weakly-supervised learning section of Fig. 1-11b, bags are illustrated with the ovals, and the instances (pixels/tiles) are shown with squares and triangles. A bag is said to be positive if it contains at least one instance coming from a WSI that is labeled positive, and it is negative if it contains only instances coming from WSIs that are labeled negative. Then,

a model can be trained that predicts a label at the bag level but which can also infer which instances contribute to the bag prediction. In other words, this means identifying which granular units (pixels or tiles) contribute the most to the bag label prediction. MIL is considered global when the goal is to identify a specific pattern in a slide (is the tissue affected by cancer or not?) while it is considered local when identification is performed at a granular level (which parts of the tile are cancerous tissue?).

3. *Unsupervised learning*: in this learning type the aim is to generate lower-dimensional representations of the input images to detect the presence of clusters or observation groups without using any preexisting labels. This is typically performed through autoencoders and their variants: these consist in an "encoder" that maps the input data into a subspace, and a "decoder" that can reconstruct the input data based on the compressed representation (Fig. 1-11b, unsupervised learning). The "bottleneck", which is in-between the encoder and the decoder, contains the feature vectors with the compressed representation of the data. These vectors can then be stacked across all observations to perform visualizations through dimensionality reduction (for example, PCA, t-SNE or UMAP).

In Section 2.3 we briefly describe the key principles of deep learning for image processing that are relevant to the problem of linking gene expression variation with patterns in histological images, which is discussed in Chapter 6.

1.2 The Genotype-Tissue Expression project

In order to set the context for the case studies in this thesis, this section describes the relevant parts of the Genotype-Tissue Expression (GTEx) project dataset, which is used in most analyses. The GTEx project was conceived in 2010 by the National Institutes of Health (NIH) [102] with the goal of building a resource for the scientific community that would aid in the characterization of human gene expression and its relationship to genetic variation and disease. Samples from up to 54 different tissues would be obtained and sequenced, producing different data types: gene expression levels, genotypes, individual medical history, and histological images with their corresponding pathology review comments. To test the feasibility of obtaining this data, GTEx first began with a 2.5-year pilot phase for sample collection through a rapid autopsy program, expecting to have around 900 post-mortem donors at the end of the project lifecycle [103]. Through the years, anonymized versions of the data have been released by the GTEx portal (see [104]), while restricted data elements such as certain individual and sample metadata are only available through data freezes in the Database of Genotypes and Phenotypes (dbGaP, see [105]).

1.2.1 Individual and sample characterization

In the final data freeze (version 8), there are 948 participating individuals (636 males and 312 females) with 17,382 RNAseq samples obtained from 54 different tissues. The tissue samples were extracted from donors in three different cohorts: post-mortem, organ donors, and surgical donors, with 515, 419 and 14 individuals, respectively. The individuals are mostly white (804), followed by black/african-american (121), with the remaining identifying under other geographically-based categories. Only adult individuals are included, with the age distribution skewing towards older ages (minimum, median and maximum age is 20, 55 and 70, respectively). Medical history for these individuals was also collected: there are 88 non-zero binary medical phenotypes, each occurring in at least one individual. The co-occurrence of these is shown in Fig. C-3. Details on biospecimen sample collection and processing are available

on the latest GTEx release [96].

1.2.2 Gene expression characterization

We define the gene expression tensor as $G \in \mathbb{R}^{n \times p \times q}$, with n = individuals, p = genes, q = tissues. In the final data freeze, $n = 948$, $p = 56{,}200$ and $q = 54$. Consider an entry G_{ijk} to be the expression level of gene j for individual i in tissue k. For a given tissue k, not all individuals $i \in \{1 \dots n\}$ are available (see Fig. C-4). However, for the available individuals, it is guaranteed that the expression of all $j = \{1 \dots p\}$ genes will be available. Muscle-Skeletal is the tissue with the largest individual sample size, while Kidney-Medulla has the lowest (803 and 4 samples, respectively; see Fig. 1-13a). G is sparse, with only 34% of the data available.

The p genes are divided into 45 categories (gene types) according to the GENCODE v26 annotation [43], with protein-coding genes being, naturally, the most abundant, followed by processed pseudogenes and lincRNAs (with 19,291, 10,141 and 7433 genes, respectively, see Fig. C-5). The expression of protein-coding genes is, in general, higher than that of lincRNAs (Fig. C-6). Although the distribution of gene expression within a single gene type is similar across tissues, each tissue has a distinctive transcriptional profile that is partially driven by tissue-specific gene expression [106]. The existence of these profiles becomes evident when reducing gene expression to a latent representation (Fig. 1-13b): by performing t-distributed Stochastic Neighbor Embedding (t-SNE, see Section 2.1.2), we observe that samples are mostly grouped together by their tissue of origin. Sample clustering is preserved when using different dimensionality reduction methods (Fig. C-7). Sample mixtures exist for tissue subtypes that belong to the same broad category, for example, skin (subtypes: sun exposed and not sun exposed), adipose tissue (subtypes: subcutaneous and visceral) and the brain tissues. This is due to these tissue pairs having closer transcriptional profiles when compared to other tissues (see Fig. C-8).

Besides tissue specificity, gene expression differences can also be observed with respect to other factors. Two examples are ischemic time and sex, which are studied in detail in Chapters 3 and 5, respectively.

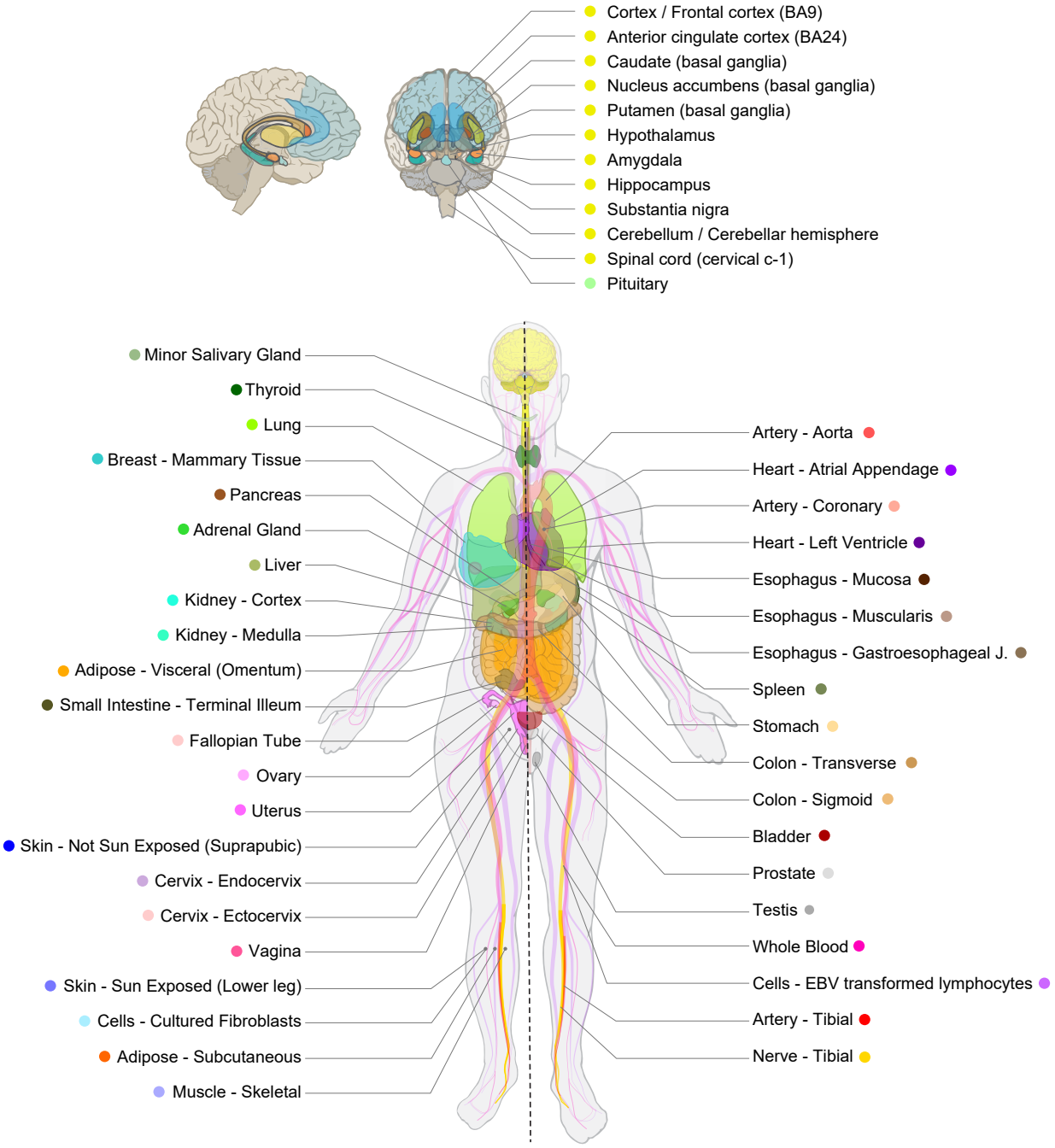


Figure 1-12: **GTEx tissue sampling sites**. Tissue sampling sites from GTEx donors. Each tissue has been assigned an unique color and abbreviation by the GTEx consortium (see Table B.1), these are used throughout the rest of this thesis work. Figure adapted from [96].

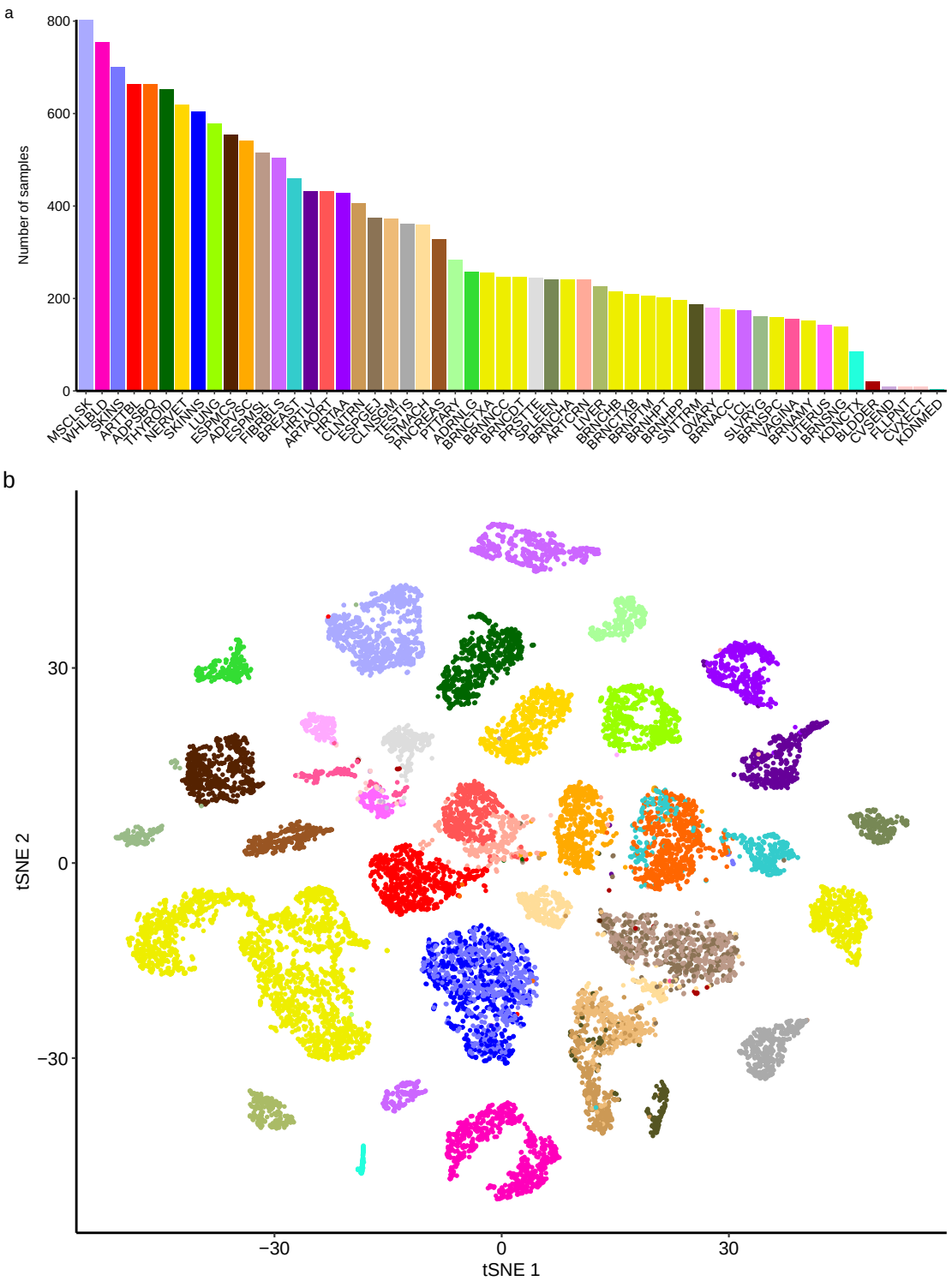


Figure 1-13: **GTEx RNAseq samples**. **(a)** RNAseq sample size per tissue. **(b)** t-SNE of 17382 RNAseq samples based on the expression of protein coding genes. Each dot corresponds to one sample.

1.2.3 Histological image characterization

Histology slides for each GTEx sample were created by the Comprehensive Biospecimen Resource (CBR), which received tissue specimens from the Biospecimen Source Sites (BSS). A high-resolution digital image of each slide was also generated by the CBR and sent to the Pathology Resource Center (PRC) for review by an American Board of Pathology-certified anatomic pathologist. As part of this assessment, the pathologists confirmed that the intented tissue morphology was present in image, identified the degree of autolysis if any, as well as the presence of other pathologic findings (such inflammation or hemorrhage). During the review process, issues such as sampling errors and image mislabeling were also resolved. Based on this review, biospecimens were deemed as "Acceptable" or "Unacceptable" for further use in the GTEx project. Additional information on the histology analysis protocol is available on the Standard Operating Procedures of the NIH's Biorepositories & Biospecimen Research Branch [107].

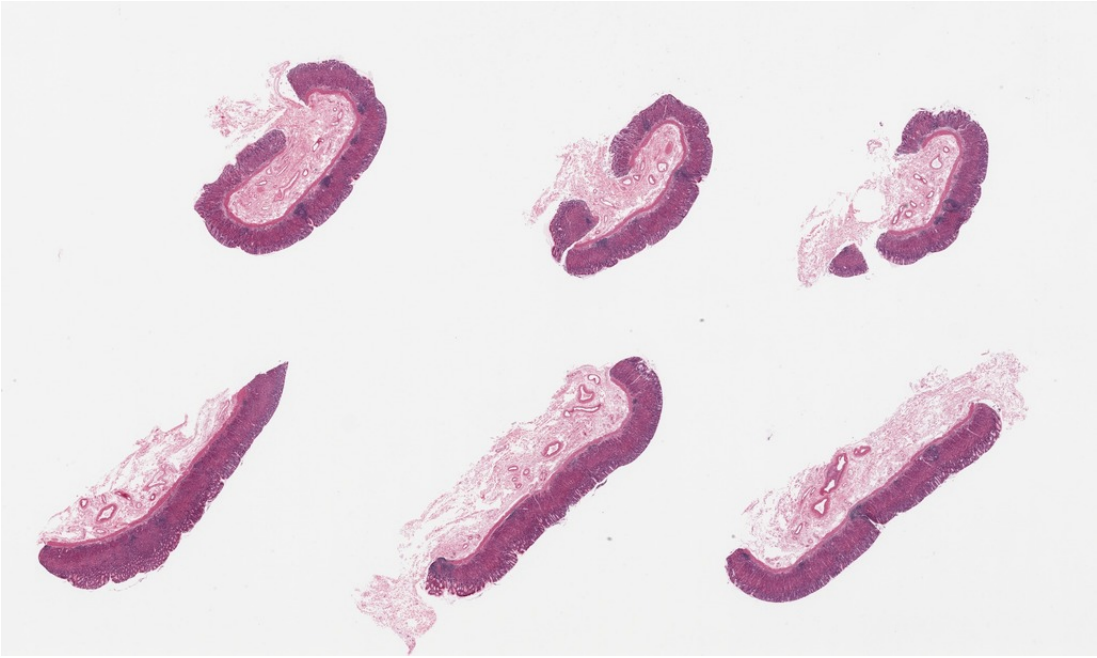
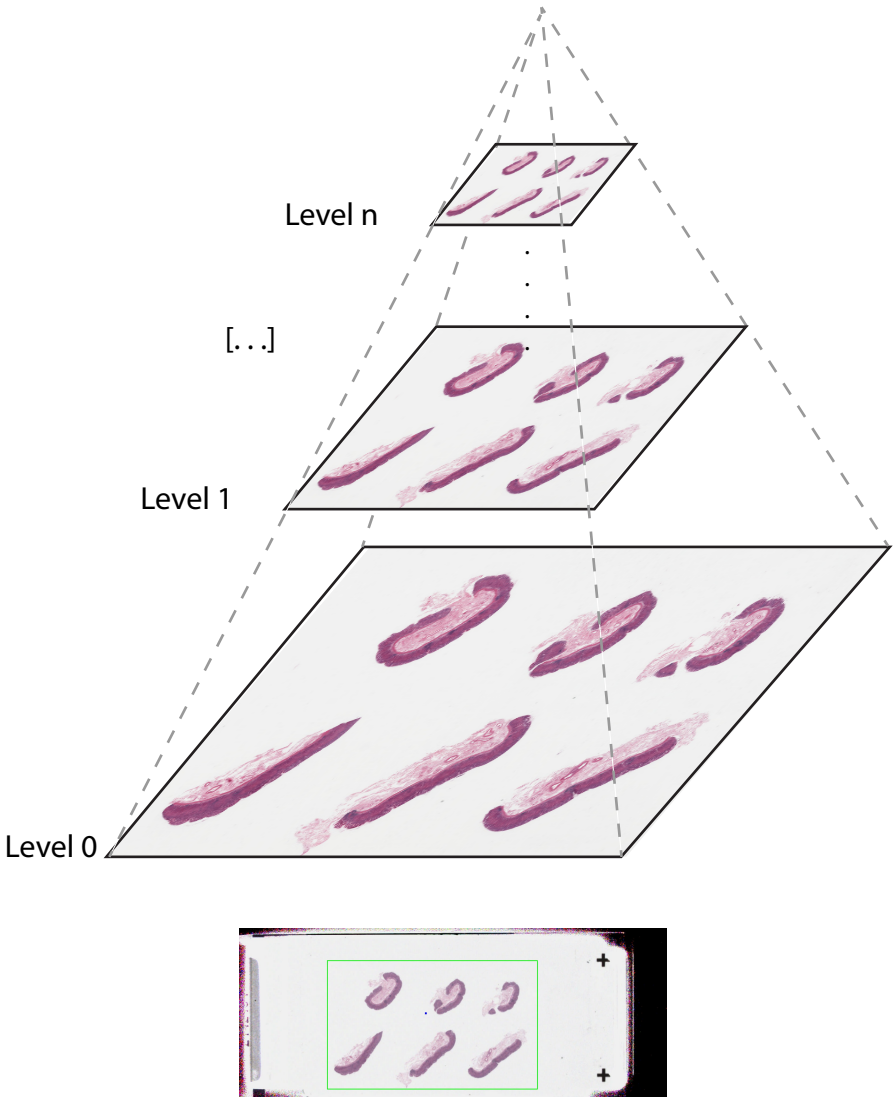


Figure 1-14: **Histological image of a stomach sample**. A downsampled version of an histological image corresponding to stomach sample GTEX-11DXX-1326. The associated pathology review comment is: "6 pieces, mild chronic active gastritis".

Each histological image may contain several sections from the specimen that was used to perform RNA-Seq (if applicable, since not all the histological images have associated sequencing data; for an image example see Fig. 1-14). These images were produced using Aperio Digital Pathology Slide Scanners with 20x objective power at 0.4942 microns-per-pixel (mpp). For the analysis performed in this thesis work, a set of 25,446 initially available histological images is considered. From these, 23,952 are deemed as acceptable, while the rest have quality issues. Only these acceptable images will be used as the starting set of images in any analysis. The largest slide has a native resolution of 109,559 $\times$ 50,518 pixels, while the smallest one has 5975 $\times$ 4743 pixels. The size distribution of these histological images will vary depending on the tissue (Fig. C-9).

Histological images are encoded into Aperio's SVS file format, which is a single-file pyramidal tiled TIFF image with associated metadata. In a SVS file, different resolutions of the same image are encoded, with each level in the pyramid containing a downsampled version of the full-resolution version of the image which is always present at level 0 (Fig. 1-15). In some SVS files, a level is reserved for an overview image of the glass slide used to obtain level 0. Loading a full-resolution image is often not possible due to large memory requirements, and depending on the application, certain techniques are necessary to be able to visualize and/or process the image. This is explored in Chapter 6.

A pathology review comment in free-text form is also associated with each histological image, as in the example shown in Fig. 1-14. These comments often contain useful histological phenotype information that can be potentially used to perform statistical association tests with other data types, such as gene expression or cell type composition. In Chapter 4, we examine how to retrieve useful keywords from these pathology review comments using Natural Language Processing (NLP) techniques.




```
Example of resolution levels encoded for stomach sample GTEX-11DXX-1326:
GTEX-11DXX-1326.svs[0] TIFF 65735x39211 65735x39211+0+0 8-bit sRGB 494.4MB
GTEX-11DXX-1326.svs[1] TIFF 1024x610 1024x610+0+0 8-bit sRGB 494.4MB
GTEX-11DXX-1326.svs[2] TIFF 16433x9802 16433x9802+0+0 8-bit sRGB 494.4MB
GTEX-11DXX-1326.svs[3] TIFF 4108x2450 4108x2450+0+0 8-bit sRGB 494.4MB
GTEX-11DXX-1326.svs[4] TIFF 2054x1225 2054x1225+0+0 8-bit sRGB 494.4MB
GTEX-11DXX-1326.svs[5] TIFF 1600x629 1600x629+0+0 8-bit sRGB 494.4MB
```

Figure 1-15: **Pyramidal structure of an SVS histological image**. Level 0 corresponds to the full resolution image of the tissue sample, while consecutive levels are downsamples of level 0 by a factor (in this case, 64x, 4x, 16x, and 32x). Level 5 is a snapshot of the glass slide (shown below the pyramid) used to obtain level 0.

1.3 From RNA to higher-order human phenotypes: thesis objectives and structure

The two main objectives of this thesis work are to:

- Relate variation in the human transcriptome with phenotypic traits through the use of statistical learning methods at the following orders of magnitude:
  - RNA: How can gene expression be used to predict a human phenotype?
  - Cell-types: How can cell-type abundance estimations be extracted from gene expression and linked to disease?
  - Tissues and organs: What is the role of cell-type composition in tissue identity?
  - Higher-order phenotypes (text and images): How can histological images from human tissues and their pathology text reports be linked to variation in gene expression?
- Determine and develop methods to encode abstract biological data (such as free-form text and images) into numerical representations.

To this end, four case studies were developed. The structure of this thesis work is as follows (see Fig. 1-16 for a graphical abstract of the topics):

Chapter 2 introduces the theoretical background and generalities for the statistical and computational methods used throughout the case studies.

Chapter 3 is an exploration on using gene expression profiles to describe a human phenotype: ischemic time, which is defined as the time elapsed since death until the preservation of a tissue sample for an individual. Using these profiles across many tissues, we construct gradient boosted tree models to predict ischemic time at the sample and individual levels.

Chapter 4 is an integrative study that describes: i) the problem of estimating cellular composition and enrichment based on gene expression profiles, ii) the extraction and normalization of biological knowledge from text annotations in such a way

that it can be used to perform statistical tests of association with molecular traits, iii) how departures from normality in cellular composition can be linked to disease, and iv) inferring high-dimensional feature vectors from histological images to validate cellular composition and disease states.

Chapter 5 presents a catalog of sex differences in gene expression with the aim of furthering the understanding of the molecular mechanisms that underlie sex-differentiated phenotypes.

Chapter 6 explores how to encode histological images to perform more complex tasks than the one explored in Chapter 4: first, we propose a computational pipeline to preprocess histological images into a usable form for downstream analyses. Second, we define a methodology to relate changes in histological patterns with variation in gene expression.

Chapter 7 is a discussion on the performed case studies, describing research lines for future work.

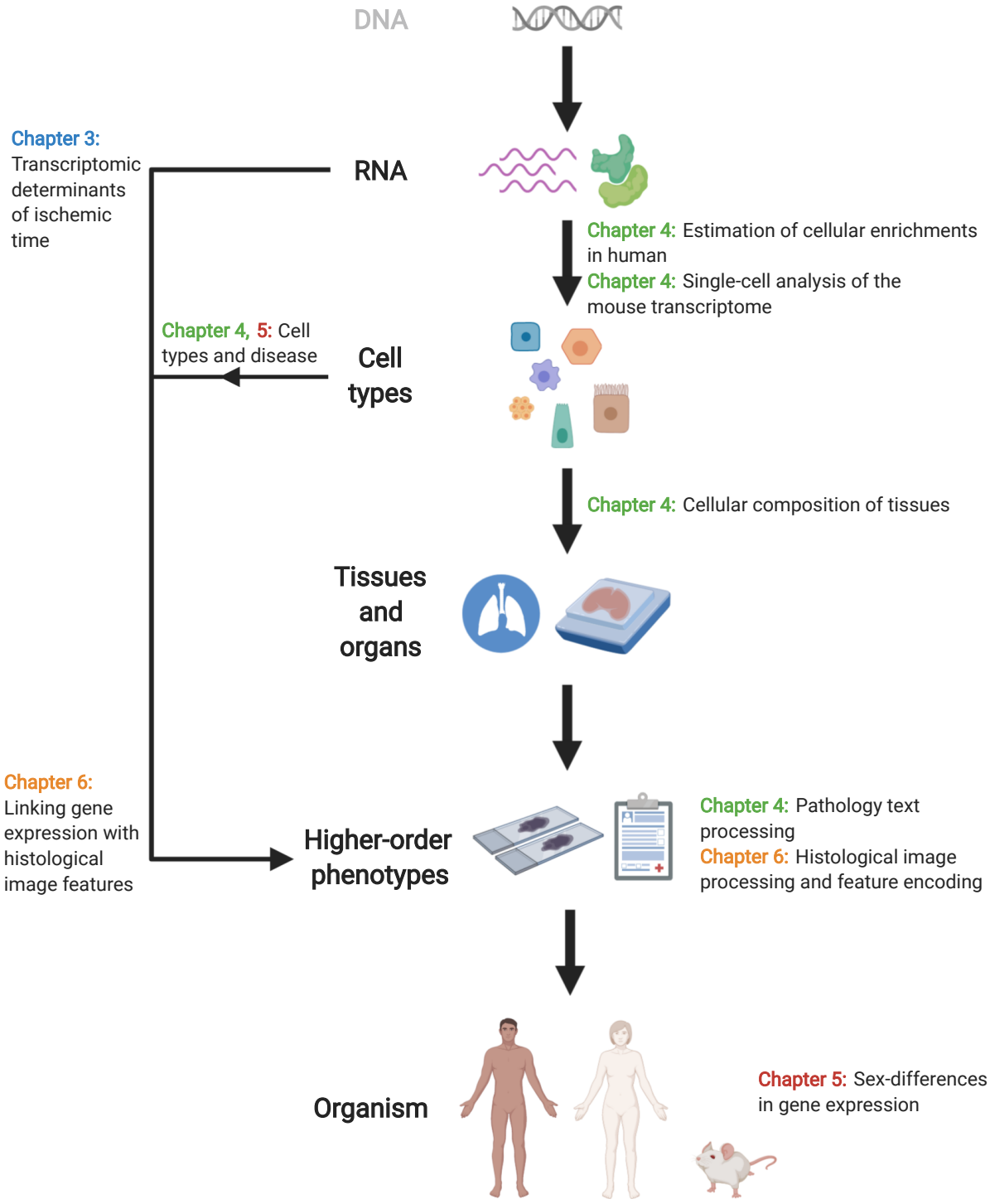


Figure 1-16: **Thesis outline.** Graphical abstract of the key topics covered in this thesis work.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Statistical learning methods in genomics

When analyzing transcriptomics datasets, the objective is usually to acquire biological insights about the data. A common task is to identify transcripts that have opposite behaviours in groups of individuals with respect to a phenotype: this is usually carried out by performing differential gene expression, which encompasses specific statistical models to test if the counts for a transcript vary across these groups. Other times, we are concerned with exploring the global structure of the data. Since it is not possible to visualize all the transcriptome dimensions at once, we use *dimensionality reduction* techniques that summarize the information contained in the data into a reduced number of latent variables. Another set of methods that fall under the umbrella term of *statistical learning* (of which dimensionality reduction is also a part of) deal with the estimation of a predictive function f that can model an outcome. Many methods to estimate f are based on machine learning algorithms, whose application in the field of transcriptomics is relatively new when compared to traditional statistical methods. Statistical learning does not only deal with prediction but also with understanding the relationship between unlabelled data observations. Along these lines, this chapter presents an exposition of different categories of methods that can have a role while performing transcriptomic studies. The specific methods described here were used in the development of the case studies in this thesis.

2.1 Dimensionality reduction

Multivariate data is pervasive in genomics, where we oftentimes aim to discover global and local structure in the data. Grasping these structures by performing simple operations like exploring the individual variances (or covariances) of the explanatory variables in the data is a task not always as intuitive as, for example, inferring knowledge from a visual representation and its associated properties. To this end, dimensionality reduction seeks to create compact representations of high-dimensional data that allow us to perform tasks such as comparing the similarity between observations or classifying them into groups. In this section, we delve into the fundamental concepts of the three different dimensionality reduction methods that are used throughout the case studies described in this thesis. Although all of these methods seek to reduce the input data into a low-dimensional embedding, each method optimizes a different objective, and thus, they lead to different maps that capture distinct aspects of the structure in the data.

2.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a classical unsupervised dimensionality reduction technique that aims to explain as much variation as possible in a dataset with a new set of uncorrelated variables called principal components. The first descriptions of PCA are considered to be given by Pearson [108] and Hotelling [109]. This technique has been widely used in biology and genomics for tasks like population structure and outlier identification in genome-wide SNP studies [110], identification of sample clusters in analyses based on gene expression [111] and cell type identification in scRNA-seq [112].

Consider a vector of random variables $\mathbf{x} \in \mathbb{R}^p$. In PCA, we find a set of weights $\boldsymbol{\alpha} \in \mathbb{R}^p$ such that a linear function $\boldsymbol{\alpha}^T\mathbf{x}$ has maximum variance:

$$\boldsymbol{\alpha}^T\mathbf{x} = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p = \sum_{j=1}^{p} \alpha_j x_j \tag{2.1}$$

We call $\boldsymbol{\alpha}_1^T\mathbf{x}$ the first principal component. Then, another set of weights $\boldsymbol{\alpha}_2$ is sought such that $\boldsymbol{\alpha}_2^T\mathbf{x}$ has maximum variance and is uncorrelated with $\boldsymbol{\alpha}_1^T\mathbf{x}$. The process is repeated for k sets of linear combinations $\boldsymbol{\alpha}_1^T\mathbf{x}, \boldsymbol{\alpha}_2^T\mathbf{x}, \ldots, \boldsymbol{\alpha}_k^T\mathbf{x}$, up to a maximum of $\min(n-1, p)$ components, where n is the number of samples in the data. Several methods exist to compute PCs, for example, the Power method, via Lagrange multipliers, the QL algorithm, and the Singular Value Decomposition (SVD) [113]. Here, we introduce the last formulation due to its ubiquity in computational software. First, we consider the sample covariance matrix $\mathbf{S} \in \mathbb{R}^{p\times p}$ of a column-centered data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with rank r:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \tag{2.2}$$

and since $\mathbf{S}$ is symmetric, it can be diagonalized as:

$$\mathbf{S} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T \tag{2.3}$$

with $\boldsymbol{\Lambda}$ being a diagonal matrix with non-increasing eigenvalues λ_i, and $\mathbf{A}$ being a matrix of eigenvectors that constitute the principal axes (or directions) of the data. The principal components (also called scores) are defined by the projection $\mathbf{XA}$ of the data onto the principal axes. The kth column of this matrix corresponds to the kth principal component. We now examine how SVD can be used to compute the PCs by decomposing the centered data matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{ULA}^T \tag{2.4}$$

where $\mathbf{U} \in \mathbb{R}^{n\times r}$ corresponds to the matrix of left singular vectors of $\mathbf{X}$, $\mathbf{A} \in \mathbb{R}^{p\times r}$ are the right singular vectors and $\mathbf{L} \in \mathbb{R}^{r\times r}$ is a positive diagonal matrix of singular values, and $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$ and $\mathbf{A}^T\mathbf{A} = \mathbf{I}_r$. By plugging (2.4) in (2.2) it becomes evident

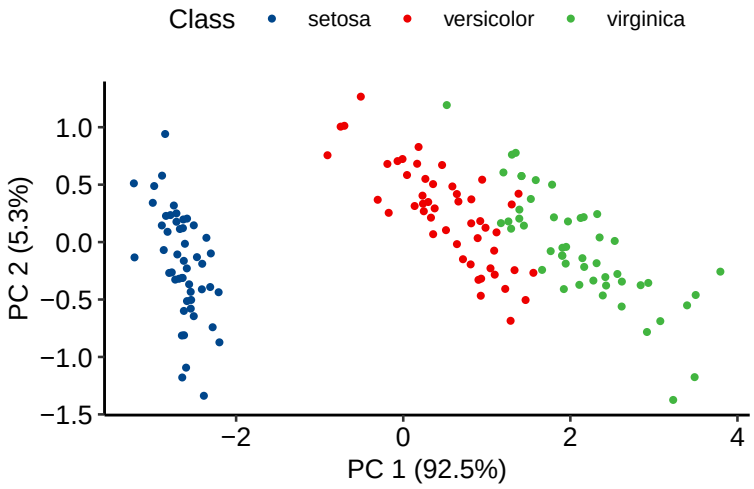


Figure 2-1: **Principal Component Analysis example.** PCA performed over the classical "iris" dataset.

that:

$$\mathbf{S} = \mathbf{A}\mathbf{L}\mathbf{U}^T\mathbf{U}\mathbf{L}\mathbf{A}^T/(n-1) \tag{2.5}$$

$$= \mathbf{A}\frac{L^2}{n-1}\mathbf{A}^\mathbf{T} \tag{2.6}$$

from which it can be seen, together with (2.3) that there is a relationship between the singular values in $\mathbf{L}$ and the eigenvalues in $\mathbf{\Lambda}$ through $\lambda_i = l_i^2/(n-1)$. Finally, it is also clear from (2.4) that $\mathbf{X}\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{A}^T\mathbf{A} = \mathbf{U}\mathbf{L}$, thus $\mathbf{U}\mathbf{L}$ are the principal components when computed through SVD.

Besides its application in many fields of biology (see, for example, in Fig. 2-1 a classical illustration of PCA with the well-known "Iris" dataset [114], and provided here only as means of comparisons with other dimensionality reduction methods, since the sample groups for this dataset are known, whereas PCA is generally used for exploratory data analysis), PCA has also found applications in the field of image processing, for example, in the task of face recognition [115]. Here we introduce an example of image manipulation and compression using PCA as it will be relevant for Chapter 6.

Consider a black and white image represented by an array of pixels $\mathbf{I} \in \mathbb{R}^{m\times q}$,

where each entry I_{ij} denotes a pixel intensity value in the interval $[0, 255]$. This can be vectorized to obtain a 1-dimensional row vector

$$\text{vec}(\mathbf{I})^T = [I_{11}, \ldots, I_{1,q}, I_{2,1}, \ldots, I_{2,q}, \ldots, I_{m,1}, \ldots, I_{m,q}] \tag{2.7}$$

A set of n images $\{\text{vec}(\mathbf{I_i})^T, \forall i \in \{1, \ldots, n\}\}$ can then be vertically stacked to obtain a matrix $\mathbf{X} \in \mathbb{R}^{n \times mq}$. Afterwards, a PCA of $\mathbf{X}$ is performed to obtain a matrix $\mathbf{T} = \mathbf{UL}$ of principal components. Notice that we can fully recover $\mathbf{X}$ with $\mathbf{TA}^T$. If we consider the case of using only a limited set of k components, we can obtain a *compressed* version $\mathbf{X}_k = \mathbf{T}_k\mathbf{A}_k^T$ of the input data $\mathbf{X}$, with reconstruction error $\left\|\mathbf{TA}^T - \mathbf{T_k}{\mathbf{A_k}}^T\right\|_2^2$.

Fig. 2-2 shows an example of image reconstruction using tiles from an histological image from a colon tissue sample, using a limited set of principal components. In this example, a set of 603 tiles (samples) of original dimensionality 224×224 is used, but resized with anti-aliasing to 96×96 only for illustrative purposes. These are vectorized into a matrix $\mathbf{X} \in \mathbb{R}^{603 \times 9216}$. Then, truncated PCA is performed to retrieve only the first 500 principal components. As expected, the reconstruction $\mathbf{X}_k$ is closer to the original when a larger number of principal components k is used. Although workarounds exist to incorporate color information with PCA (for example, by vectorizing the 3-way array used to represent a color image), in practice, PCA is a limited option to perform complex image processing tasks since the relation between pixels is lost when performing vectorization of an image, in addition to complications related to wide variations in the images. For these reasons, it tends to work well only over uniform datasets with few variations in positioning and lighting.

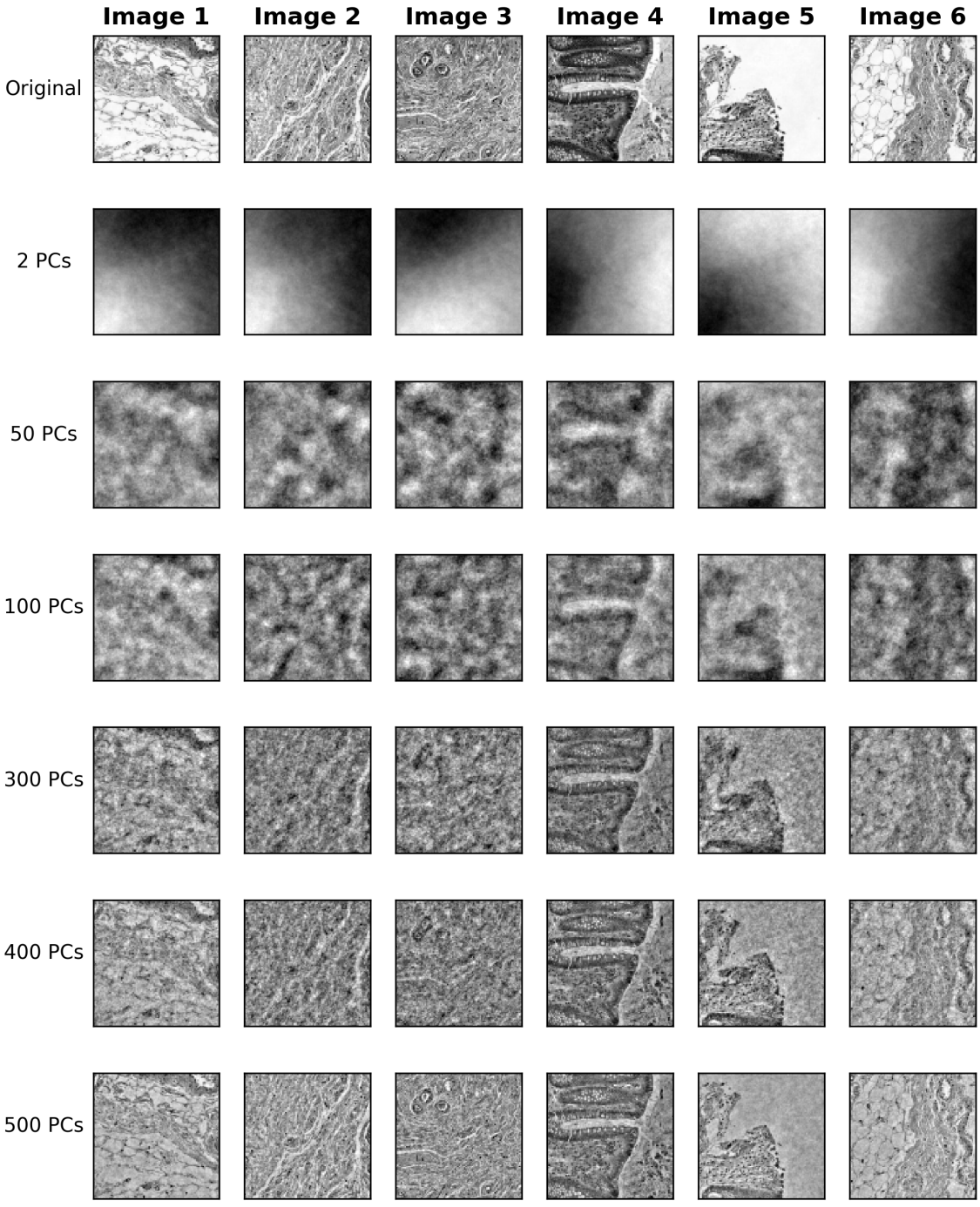


Figure 2-2: **Black and white image reconstruction with PCA.** Six histology image tiles from a GTEx colon sample (first row, corresponding to tiles of sample GTEX-11DXX-1825) are reconstructed with a given number of PCs, shown on the left side on each of the rows.

2.1.2 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised data visualization technique that attempts to capture local patterns in the data as well as reveal the presence of clusters at larger scales. It was introduced in 2008 by van Der Maaten and Hinton [116] and since then, it has been used in genomics to perform unsupervised analysis of gene expression, especially in the context of single-cell RNA sequencing (scRNA-seq) which can range from a few thousands up to millions of cells (observations). t-SNE is also commonly applied in other tasks within biology since it tends to yield good visualizations in datasets with a hierarchical organization among the observations [117].

Consider a set of N high-dimensional vectors $X = \{\mathbf{x}_1, \dots \mathbf{x}_N\}$. t-SNE aims to construct a low-dimensional map $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ which usually is two or three-dimensional. First, the similarity between the points in X is computed as a joint probability distribution P. Each entry in the matrix P is:

$$p_{ij} = \frac{\exp(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / \sigma)}{\sum_k \sum_{l \neq k} \exp(-\left\|\mathbf{x}_k - \mathbf{x}_l\right\|^2 / \sigma)}, \forall i \forall j : i \neq j \tag{2.8}$$

which is defined only only over pairs of non-identical points. Thus, $p_{ii} = 0$ since only pairwise similarities between points are of interest. In the low-dimensional map Y, the similarity is measured with a joint distribution Q, with pairwise distance q_{ij} between the observations first defined as:

$$q_{ij} = \frac{\exp(-\left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2)}{\sum_k \sum_{l \neq k} \exp(-\left\|\mathbf{y}_k - \mathbf{y}_l\right\|^2)}, \forall i \forall j : i \neq j \tag{2.9}$$

and also setting $q_{ii} = 0$, just as in the case of the pairwise similarities over X. This formulation for q_{ij} was introduced in Stochastic Neighbor Embedding (SNE) [118]. Then, the objective function of t-SNE seeks to minimize the Kullback-Leibler divergence between P and Q:

$$\min C(Y) = \min \mathrm{KL}(P \| \| Q) = \min \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.10}$$

However, when minimizing this objective using Q as defined by the pairwise similarities in (2.9), the data points will suffer from a crowding problem related to the volume differences between X and Y: if X is high-dimensional, pairs of points that are slightly similar would be represented as far away in Y, and since this happens for many pairs of points due to the high volume of X, the points will be squashed and "crowded" together in Y [119]. The effects of dimensionality on volume have already been discussed in Section 1.1.3. To alleviate this, q_{ij} is instead redefined to be proportional to a Student-t distribution with one degree of freedom:

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}, \forall i \forall j : i \neq j \tag{2.11}$$

The distribution of Q has heavier tails and thus allows to represent larger distances in the Y that correspond to only moderate distances in X. Minimization of the objective function (2.10) is performed through gradient descent, with the gradient update given by:

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}) \tag{2.12}$$

with $\mathcal{Y}^{(t)}$ being the solution at iteration t, η is the learning rate, and $\alpha(t)$ is the momentum. Initialization of $\mathcal{Y}^{(0)}$ is done by random sampling from a Gaussian with small variance and centered at the origin. The gradient of the cost function is given by:

$$\frac{\delta C}{\delta \mathbf{y}_i} = 4 \sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} \tag{2.13}$$

In t-SNE, it is also necessary to fix a parameter called *perplexity*, which is used as a proxy to choose values of σ in (2.8) and has the effect of balancing the objective towards recovering local or global structure in the data. This value can be roughly interpreted as the number of neighbors for each data point and may need to be tweaked depending on the size of the data. We defer to [116] for its formal definition as well as for the derivation of (2.13). Choosing a good value for perplexity is important when using t-SNE since it tends to have a larger effect over the final map (see, for example, Fig. 2-3) when compared, for example, with the number of iterations for gradient

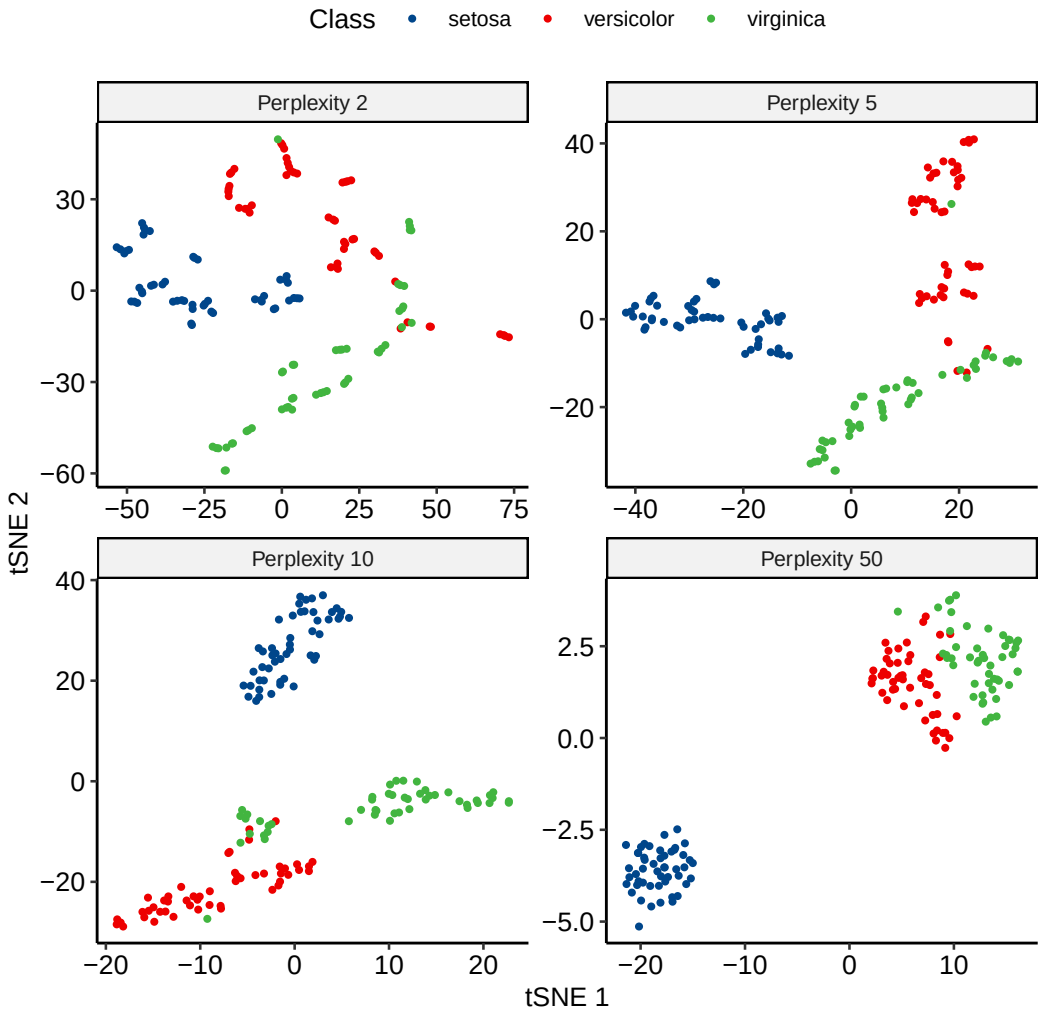


Figure 2-3: **Effect of perplexity in t-SNE.** t-SNE performed over the classical "iris" dataset with different values of perplexity.

descent (as long as the map is stable enough) [120].

Due to the nature of the joint distribution Q, care must be exercised when interpreting t-SNE maps. For example, the distances between the clusters might not necessarily have a meaning: this can be seen, for example, in Fig. 2-3 where the distance between species clusters is conditioned on the attention given to the global/local structure by the chosen perplexity value. A detailed exploration of the nuances of t-SNE is presented in [120].

2.1.3 Uniform Manifold Approximation and Projection

Recently, non-linear dimensionality reduction methods such as t-SNE have been used to visualize the global structure of single-cell gene expression and reveal cell type populations. These types of methods help avoid the overcrowding that happens when clusters of data points are located in an overlapping areas [121], an issue that tends to happen, for example, in PCA. Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction method that was recently introduced and that is already being used to perform single-cell analyses due to its ability to preserve both local and global structures in the data. Its derivation is based on concepts from algebraic topology, manifold learning and Riemmanian geometry. Here, we will present the main definitions of UMAP as described in the original work [122].

The summary of the mathematical idea of UMAP is to consider k-dimensional simplices (convex hull of $k+1$ data points) to create simplicial complexes that describe the underlying topological space of the data, given a sample. To do this, an open cover (sets whose union represent the whole space) of the topological space is generated and a Čech complex (a type of simplicial complex) is built from this. Since in practice we work with finite samples, the open cover is generated through balls of a fixed radius (through some distance, since the sample is assumed to be in a metric space) around a point (see equation 2.15 for a clearer notion). UMAP then tries to generate a low-dimensional embedding that that preserves the topology of the input space, based mainly on 0- and 1-simplices.

From a computational perspective, this process is approximated with two steps: constructing a graph of the data points (the proxy for the fuzzy simplicial sets) and optimizing the graph layout in such a way that the topological structure of the data is preserved. First, from a dataset $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and a distance measure $d : X \times X \to \mathbb{R}_{[0,\infty)}$, the k-nearest neighbors $\{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}\}$ are computed for each data point $\mathbf{x}_i$. For each point, the distance ρ_i to the nearest neighbor is also computed:

$$\rho_i = \min\{d(\mathbf{x}_i, \mathbf{x}_{i_j}) \mid 1 \leq j \leq k, d(\mathbf{x}_i, \mathbf{x}_{i_j}) \geq 0\} \tag{2.14}$$

as well as the neighborhood σ_i of $\mathbf{x}_i$, chosen so that:

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(\mathbf{x}_i, \mathbf{x}_{i_j}) - \rho_i)}{\sigma_i}\right) = \log_2(k) \tag{2.15}$$

With these, we can generate a weighted graph $\bar{G} = (V, E, w)$ with the vertices being the data points, the edges being the sets of k-nearest neighbors for all data points and the weights computed between each point and its neighbors:

$$w_i(\mathbf{x}_i, \mathbf{x}_{i_j}) = \exp\left(\frac{-\max(0, d(\mathbf{x}_i, \mathbf{x}_{i_j}) - \rho_i)}{\sigma_i}\right) \tag{2.16}$$

$$w_{i,i_j} = w_i(\mathbf{x}_i, \mathbf{x}_{i_j}) + w_{i_j}(\mathbf{x}_{i_j}, \mathbf{x}_i) + w_i(\mathbf{x}_i, \mathbf{x}_{i_j}) w_{i_j}(\mathbf{x}_{i_j}, \mathbf{x}_i) \tag{2.17}$$

Since $w_i(\mathbf{x}_i, \mathbf{x}_{i_j})$ is asymmetrical, the actual weights are just symmetrized in the second line of the equation above. A weight matrix w' is also computed for the data points in the low-dimensional embedding Y, using an approximation for membership strength based on two hyperparameters a and b:

$$w'_{ij} = \frac{1}{1 + a d(\mathbf{y}_i, \mathbf{y}_j)^b} \tag{2.18}$$

We now have everything to define UMAP's cost function C, which is the cross-entropy of the two fuzzy simplicial sets:

$$C(w, w') = \sum_{ij} \left[w_{ij} \log\left(\frac{w_{ij}}{w'_{ij}}\right) + (1 - w_{ij}) \log\left(\frac{1 - w_{ij}}{1 - w'_{ij}}\right) \right] \tag{2.19}$$

With this objective, we find points $\{\mathbf{y}_i, \ldots, \mathbf{y}_n\}$ whose weighted graph H (low-dimensional representation) approximates the graph G (input space). This can be thought of as a non-convex optimization problem that is similar to optimizing a force-directed graph layout, with the first term of (2.19) modeling the attraction between the points and the second controlling the repulsion. It can be optimized through sampling-based gradient descent (the derivatives are easy to compute), and initialized, for example, with a spectral embedding. An example of UMAP is shown in Fig. 2-4, where the

variability of the embedding is shown with respect to the number of neighbors k which balance local versus global structure, and the minimum distance between the points in the embedding.

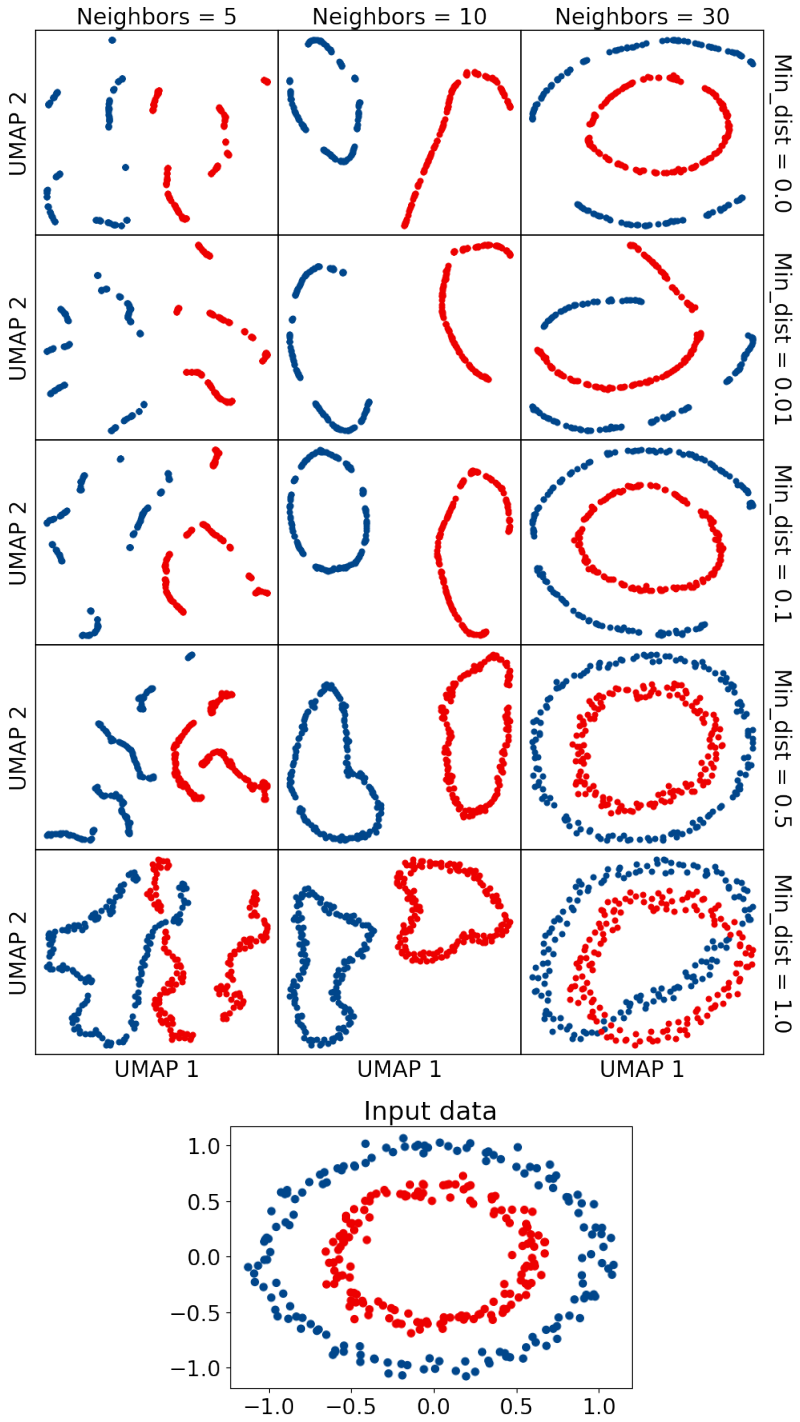


Figure 2-4: **UMAP parameters.** UMAP representations with varying minimal distance and number of neighbors. The original input data is shown at the bottom.

2.2 Statistical learning

The term *statistical learning* refers to computational methods and algorithms that aid in understanding data [123] that can be either labeled or unlabelled. In the first case, we speak of *supervised learning* since generally the goal is to learn a predictive function to estimate an output based on a set on inputs; types of supervised learning tasks include classification, regression and ranking. The second case is usually addressed through *unsupervised learning* methods that aim to identify the relationships between the observations in a dataset; for example, through clustering. These are two of the main learning paradigms but others exist, such as semi-supervised learning where data can be weakly-annotated or with labels existing only partially, and reinforcement learning which is based on agents trying to maximize a reward by taking actions in a controlled environment.

In a practical sense, the process of learning from data is streamlined through machine learning, and this is one of the reasons why these methods have become commonplace in the field of bioinformatics where the need exists to understand large volumes of data and produce inferences from it. On the other hand, statistical learning theory aims to formalize how that learning process occurs. In Section 1.1.3, a glimpse was provided on how to measure learning when building a predictive function. When learning from data, it is possible to build a function that will perfectly estimate that data, however, this is not of interest: what we seek is for the predictive function to pick up patterns in the training data that can generalize to unseen observations. The ability to generalize is related to the complexity of the model, since ideally we would want a model that is simple but that is also good at fitting the data.

Measuring the complexity of a model is not exactly straightforward: for example, in parametric statistical modelling, complexity could be thought of as the number of parameters in the model. In machine learning, we usually work with infinite-dimensional hypothesis sets $\mathcal{H}$, but concepts from statistical learning theory such as the Rademacher complexity or the Vapnik-Chervonenkis (VC) dimension provide guarantees about the learning capabilities for $\mathcal{H}$ from finite samples, although through

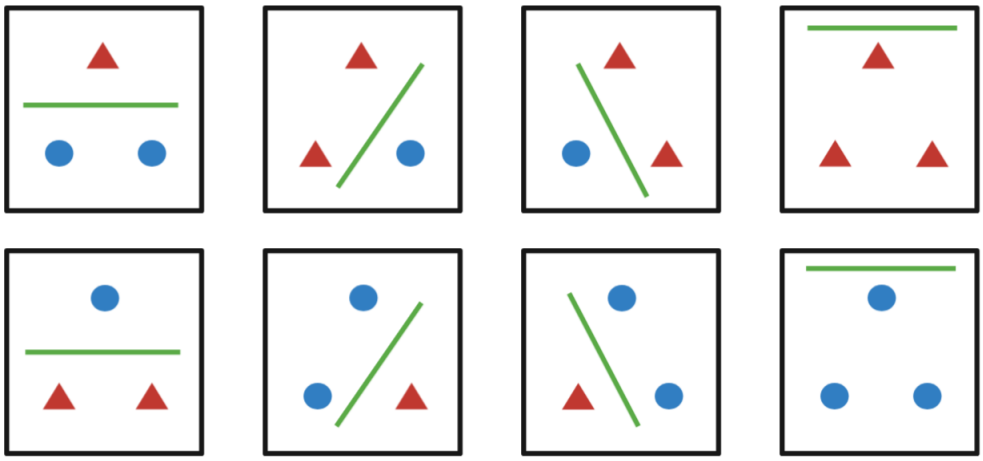


Figure 2-5: **VC dimension of a classifier.** The VC dimension for a model f based on a straight line is 3: at most three points can be perfectly separated in all possible class arrangements in the data.

different conceptualizations of complexity [124]. To understand this intuitively, we can illustrate what VC dimension is: consider a set of data points of cardinality S, then, the VC dimension of a binary classifier model f is the maximal cardinal S for which all the points can be shattered by f: in other words, if a set of parameters θ exists such that f can perfectly separate a set of at most S points in any possible label arrangement, then the VC dimension is S. In Fig. 2-5, we consider a model f that is simply a straight line and a set of 3 points with two classes. 2^3 possible arrangements exist, and f can perfectly separate all the cases. However, if we consider a set of 4 or more, we can clearly see that this is not possible, since not all arrangements can be separated. Thus, the VC dimension of f is 3.

The fact that no unique way exists to measure complexity is related to the No Free Lunch theorem [125], which, in short, states that there does not exist a single model that will perform the best in all situations in the sense that if we take the average performance of two different algorithms evaluated over a large set of different problems, we will see that the performance is identical, and this happens because each model makes a different set of assumptions about the data. We can, however, control within-model complexity through the *bias-variance* tradeoff. If we assume data generated through a model $Y = f(X) + \epsilon$ for which $E(\epsilon) = 0$, we can decompose

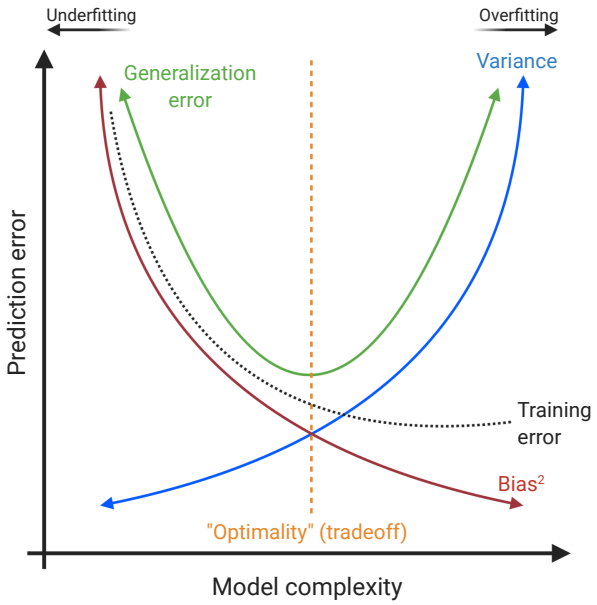


Figure 2-6: **Bias-variance tradeoff.** Behavior of bias and variance with respect to model complexity and prediction error.

the expected mean square error at a point x in three terms: variance of $\hat{f}(x)$, bias of $\hat{f}(x)$ and the irreducible error $\text{Var}(\epsilon)$:

$$E(y - \hat{f}(x))^2 = \text{Var}(\hat{f}(x)) + [\text{Bias}(\hat{f}(x))]^2 + \text{Var}(\epsilon) \tag{2.20}$$

which refers to the error we would expect if we evaluated x when estimating f repeatedly with many different sets of training data. Through the parameters of the model, we can decide if we want to compromise variance, which refers to the variability in f as a function of the training data, or favor bias, which refers to the errors that are inherent to simplifying the problem. Finding a good tradeoff between both quantities (see Fig. 2-6) is how we avoid underfitting and/or overfitting a model.

In this section we cover four methods that although not specific to computational biology, can be used to understand biological datasets. Some of these, such as hierarchical clustering, have been extensively used in the field, while others, such as gradient boosting and model explanation frameworks based on Shapley values (the latter adapted from game theory concepts) are not as commonplace but have found recently found applications.

2.2.1 Hierarchical clustering

Clustering is a type of unsupervised analysis performed when the goal is to identify which observations of a dataset tend to be closely related to each other through a quantification of their degree of (dis)similarity. Observations that belong to the same cluster tend to share properties among themselves as opposed to observations assigned to different clusters. How the observations are grouped strongly depends on the dissimilarity measure used to compare the observations, as well as on the clustering algorithm. Many clustering algorithms depend on a matrix of pairwise dissimilarities $\mathbf{D} \in \mathbb{R}^{n \times n}$, with n referring to the number of data points x_i with p variables in a dataset $\mathbf{X} \in \mathbb{R}^{n \times p}$. Through a function $d_j(x_{ij}, x_{i'j})$ definining a dissimilarity between attribute j in both observations (for example, euclidean distance for the case of continuous values), we can define a dissimilarity at the level of the observation. For example, for a pair of objects x_i and $x_{i'}$:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j}) \tag{2.21}$$

With this, we evaluate all pairwise combinations of objects to generate $\mathbf{D}$, which is assumed to be symmetrical for most algorithms:

$$\mathbf{D}_{ii'} = \mathbf{D}_{i'i} = D(x_i, x_{i'}) = D(x_{i'}, x_i) \tag{2.22}$$

Here, we introduce the generalities of hierarchical clustering, which is a type of clustering analysis that bundles observations through a hierarchical arrangement. This procedure does not require setting the number of clusters a priori, like in the case of k-means. Instead, hierarchical clustering can be performed through two different conceptualizations: agglomerative (bottom-up) and divisive (top-down) [47]. These refer to the way that the clusters are defined: agglomerative starts by considering each observation as a singleton cluster and then merges a pair of clusters according to a specific criterion that seeks to minimize the distance between clusters, reducing the number of clusters by 1 in each step until the root of the hierarchy is only one

cluster (see Algorithm 1). For an example of agglomerative hierarchical clustering, see Fig. C-8.

Algorithm 1: AGGLOMERATIVE CLUSTERING

Data: $\mathbf{D} \in \mathbb{R}^{n \times n}$

Result: Dendrogram

Initialize: $C = \{C_1 = \{x_1\}, \ldots, C_n = \{x_n\}\}$

while $|C| \neq 1$ **do**

  $(C_k, C_{k'}) = \text{argmin}_{A,B} f_{\text{SINGLE}}(C_A, C_B),\ A, B \in C$

  $C = C \setminus \{C_k, C_{k'}\}$

  $C = C \cup \{\{C_k, C_{k'}\}\}$

end

Examples of dissimilarity measures between two clusters C_A and C_B include single linkage, which considers the distance between the two clusters to be the pair of instances with the smallest distance:

$$f_{\text{SINGLE}}(C_A, C_B) = \min_{\substack{i \in C_A \\ i' \in C_B}} D(x_i, x_{i'}) \tag{2.23}$$

and complete linkage, which takes the inter-cluster distance to be the most dissimilar pair (maximal distances) of instances:

$$f_{\text{COMPLETE}}(C_A, C_B) = \max_{\substack{i \in C_A \\ i' \in C_B}} D(x_i, x_{i'}) \tag{2.24}$$

Measures such as average linkage or Ward's criterion are also common. A divisive strategy will start from the top of the hierarchy, with all data points grouped into a single cluster, and creating subsequent partitions between pairs of clusters with the largest distance according to a specific criterion. Both strategies are generally monotonic, in the sense that dissimilarity between clusters will always increase with respect to the cluster split level. Thus, cluster pairings can be represented through a dendrogram. The intepretability conferred by dendrograms have popularized their use in bioinformatics.

2.2.2 Gradient boosting

Machine learning classification methods that are able to robustly deal with high dimensional spaces have allowed researchers to build models that use different types of -omics datasets to distinguish between different groups of observations with respect to phenotypic traits, or with disease [126]. One such method is *gradient boosted trees*, which has been widely used in the last years due to its overall good performance in different types of tasks [127] that rely on tabular data. Here, a common conceptualization of gradient boosting trees called *XGBoost* will be described, following the exposition by Chen and Guestrin [128].

Consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ with number of examples $|\mathcal{D}| = n$ and data dimensionality $\mathbf{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}$. For any task such as regression, classification and ranking, we seek to find an optimal set of parameters θ obtained by training a model that minimizes an objective function that measures how well predictions $\hat{y}_i = f(\mathbf{x}_i)$ approximate y_i. To this end, a regularized learning objective can be formalized as:

$$\mathcal{L}(\theta) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2.25}$$

with l being a differentiable convex loss function and Ω being a regularization term that penalizes the complexity of the model. In the context of tree boosting, the output y_i for a data point is predicted with K additive functions:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F} \tag{2.26}$$

with $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}$ being the space of possible regression trees, and q being a tree structure which maps a data point to leaf in that tree: $q : \mathbb{R}^p \to \{1, \ldots, T\}, w \in \mathbb{R}^T$, where T is the total number of leaves in a tree. Notice that here, each f_k is independent, meaning that it has its own set of leaf weights w and tree structure q. With this, from (2.26) we realize that the final prediction for a sample is given by the sum of the corresponding leaf scores across all trees. With this context established,

we can set the objective penalty as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{2.27}$$

which aims to prevent overfitting by penalizing the leaf weights, and thus leading to simpler models, which in the case of other tree learning methods was dealt with by applying heuristics focused on improving impurity. It becomes evident that since (2.25) contains functions as parameters, we cannot directly use stochastic gradient descent to minimize this function, and thus, we rely on additive training (boosting) to optimize the objective. At the t-th iteration, with prediction $\hat{y}_i^{(t)}$ of the i-th instance, we add f_t to greedily improve the model:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \tag{2.28}$$

The objective can be optimized using first and second order Taylor approximations $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$:

$$\mathcal{L}^{(t)} \simeq \left[\sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)\right] + \Omega(f_t) \tag{2.29}$$

and further simplified by removing the constant term $l(y_i, \hat{y}_i^{(t-1)})$:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)\right] + \Omega(f_t) \tag{2.30}$$

Note that here, the objective depends on g_i and h_i. Therefore, custom loss functions can be used as long as g_i and h_i are known: this is how *XGBoost* can deal with other types of problems besides classification: regression, ranking, etc. To measure how good a tree structure $q(\mathbf{x})$ is, we reformulate (2.30) by i) expressing the trees $f_t(\cdot)$

with their leaf scores $w_{q(\cdot)}$, ii) explicitly stating the regularization term (2.27):

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i w_{q(\mathbf{x}_i)} + \frac{1}{2} h_i w_{q(\mathbf{x}_i)}^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{2.31}$$

$$= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{2.32}$$

with $I_j = \{i | q(\mathbf{x}_i) = j\}$ being the indices of the observations pushed into leaf j. We are able to rewrite (2.31) in terms of the tree indices in (2.32) because the observations that belong to the same leaf have the same score. Now, let's assign terms to the summations: $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$. We can now rewrite as:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} \left(H_j + \lambda \right) w_j^2 \right] + \gamma T \tag{2.33}$$

Notice here that the summation term is quadratic, and the best w_j for a given $q(\mathbf{x})$ will be:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{2.34}$$

By plugging this into (2.33), we finally obtain an equation that provides a measure of the goodness of the tree structure $q(x)$:

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{2.35}$$

We defer to [128] for algorithmic details on how to iteratively add branches to the tree in order to evaluate split candidates. In practice, it is important to perform parameter tuning (for example, via grid search + cross-validation, or Bayesian optimization) to find a good bias-variance tradeoff point. If a large number of estimators (trees) is used, the decision boundary (in the case of classification) will overfit the data, not being able to generalize (see, for example, Fig. 2-7 with 30 estimators).

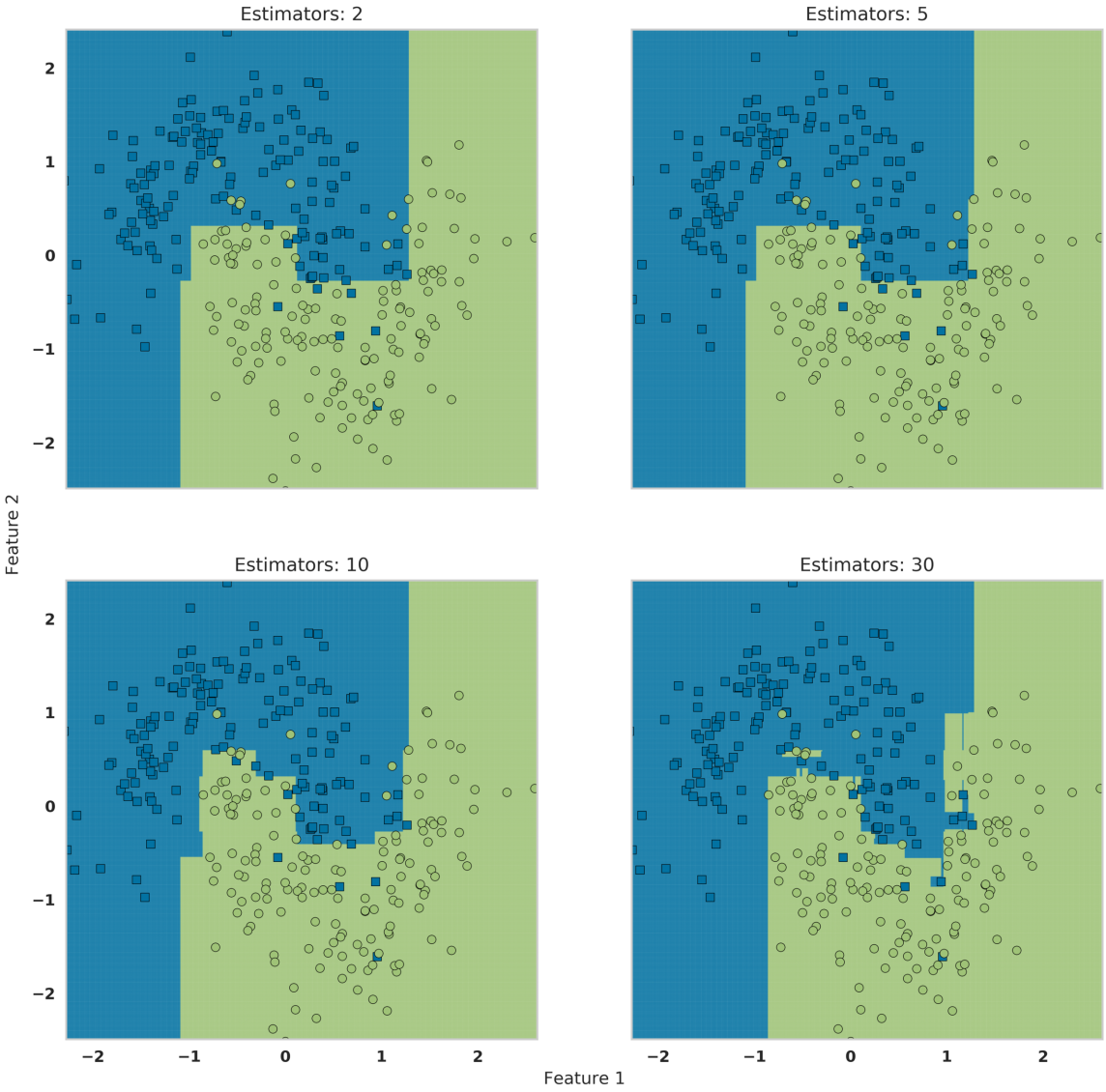


Figure 2-7: **Decision boundaries in gradient boosted trees.** Here we show the effect of the number of estimators on the decision boundary for a toy classification problem with two features and two classes. The granularity of the decision boundary will increase as the number of estimators increases. A large number of trees without proper regularization will lead to overfitting, as shown here with 30 estimators.

2.2.3 Hyperparameter search with Bayesian optimization

Many machine learning methods depend on a set of hyperparameters λ to control model complexity and make a tradeoff between bias and variance, as discussed earlier in Section 2.2. These hyperparameters are different from those used as data weights when fitting the predictive function, since the former are inherent to the algoritm used for the learning task, where we ultimately seek to minimize a loss function $\mathcal{L}(\mathbf{X}_{\text{test}};\ f)$ for a model f over a test dataset. The model f is fit based on an algorithm $\mathcal{A}$ which depends on the hyperparameters λ found to be optimal in a training dataset $\mathbf{X}_{\text{train}}$: $f = \mathcal{A}(\mathbf{X}_{\text{train}};\ \lambda)$. Thus, hyperparameter search refers to the process of finding a set λ^* that produces an optimal model f^* [129]:

$$\lambda^* = \underset{\lambda}{\text{argmin}}\, \mathcal{L}(\mathbf{X}_{\text{test}};\ \mathcal{A}(\mathbf{X}_{\text{train}};\ \lambda)) \tag{2.36}$$

Sets of candidate hyperparameters can be derived, for example, through random search in a space bounded between reasonable minimal and maximal values for each hyperparameter, or also through grid search in which candidate values for each hyperparameter are proposed through a specific sampling criteria, for example, by selecting a number of equidistantly spaced points between the lower and upper bounds for that hyperparameter. It becomes evident that creating a granular grid (or even random search) for many parameters can become prohibitive in terms of the number of evaluation needed to find λ^*, especially within the context of procedures that aim to describe model robustness (through the inspection of the error criterion, for example) such as bootstrapping or cross-validation.

Bayesian optimization provides a framework that can be used to make educated guesses for λ^* candidates that tend to behave better than baseline models or random search, and that, although requiring additional computations, can reduce overall training time since less evaluations are needed to find an optimal point. Here, we very briefly touch the practical considerations for this framework, as presented in [130], and that are based on two main elements. The first is the choice of a prior distribution (also called a surrogate model) over functions $f : \mathcal{X} \to \mathbb{R}$ that model the assumptions

of the function we need to optimize: Gaussian processes (GP) allow this by inducing a multivariate Gaussian distribution on a set of N points $\{\mathbf{x}_n \in \mathcal{X}\}_{n=1}^{N}$. We assume then that $f(\mathbf{x})$ is drawn from a GP, with each observation having a label $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$ that follows a normal distribution $y_n \sim \mathcal{N}(f(\mathbf{x}_n, \nu))$ with ν being the noise variance of the observations. The second consideration is the derivation of an utility function based on the posterior distribution from which we can decide where to sample next, based on a criterion of expected loss: the *acquisition* function $a : \mathcal{X} \to \mathbb{R}^+$:

$$\mathbf{x}_\mathrm{t} = \underset{\mathbf{x}}{\operatorname{argmax}}\, a(\mathbf{x}|\{\{\mathbf{x}_1, y_1\}, \ldots, \{\mathbf{x}_{t-1}, y_{t-1}\}\}) \tag{2.37}$$

with $\mathbf{x}_t$ being the next sampling point obtained by evaluating the acquisition function for the current data point conditioned over all previously drawn points. Once the sampling point has been obtained, we estimate the response $y_t = f(\mathbf{x}_t) + \epsilon_t$, add the pair $\{\mathbf{x}_t, y_t\}$ to the list of sampled points, and then repeat (acquire a new sampling point). The *expected improvement* acquisition function has been shown to perform better than other classical acquisition functions such as the probability of improvement [130]. As its name suggests, it can be defined as the expectation of the maximal difference between evaluating a sample $\mathbf{x}$ and the current best point $\mathbf{x}^+ = \operatorname{argmax}_{\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}} f(\mathbf{x}_i)$:

$$a_\mathrm{EI} = \mathbb{E}[\max(f(\mathbf{x}) - f(\mathbf{x}^+), 0)] \tag{2.38}$$

We refer to [131] for details on how to analytically evaluate the GP. In the case of hyperparameter optimization, the sampling points correspond to candidate hyperparameters for λ^* and the objective function is defined by the machine learning problem (for example, MSE in the case of regression, negative log-likelihood in classification, etc.).

Fig. 2-8 shows an example of maximizing a known noise-free target function (in red, and which is unknown in practice) through a GP. The left panel in the first row shows two initial guesses that that generate an unhelpful posterior (dashed black lines) but that will be the starting point for the optimization. We then start sampling

candidate points in each step of the GP, and see the next candidate point to sample according to the utility function (panels on the right). As we sample more points, the uncertainty decreases. In Chapter 5 we perform Bayesian optimization for candidate hyperparameter search while building classifier models.

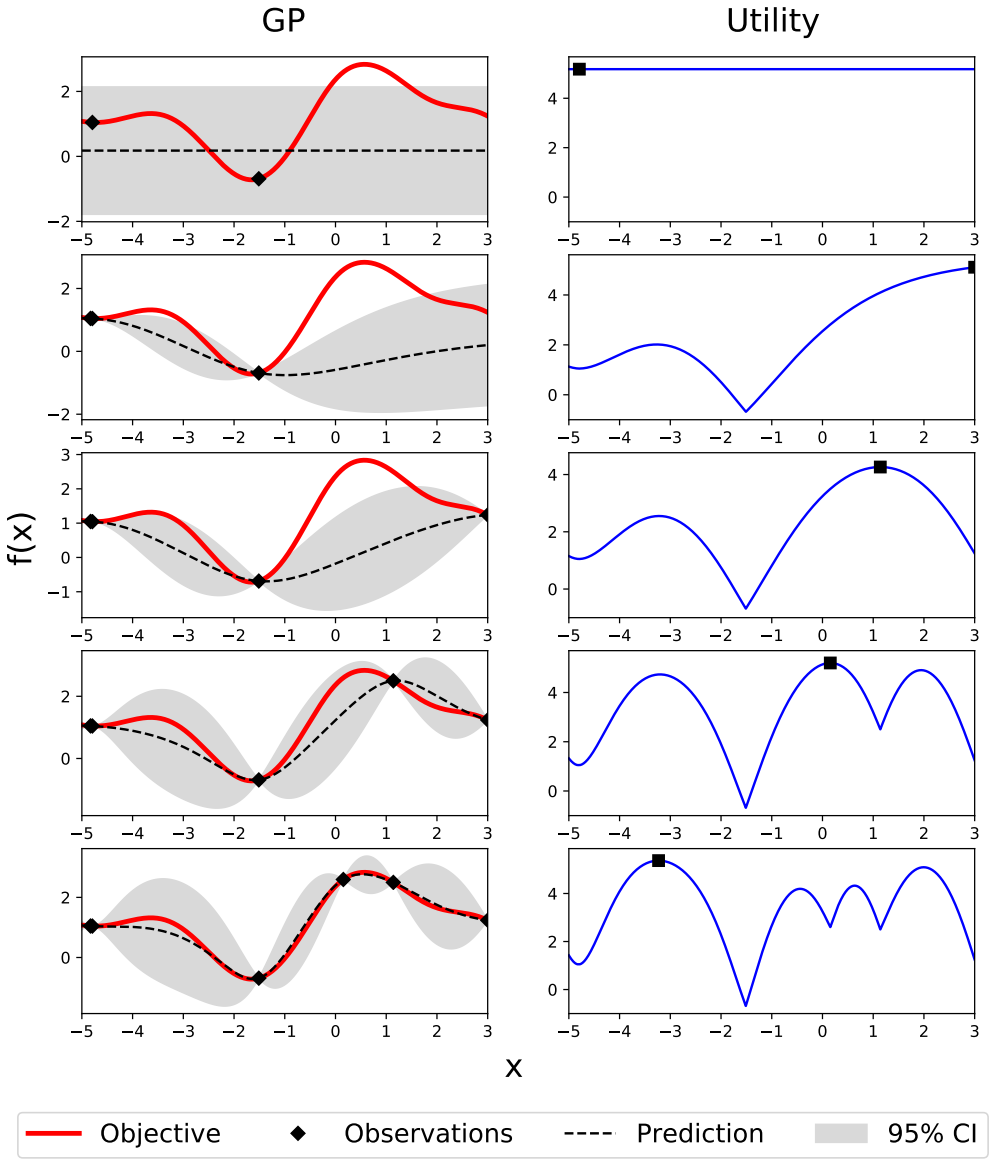


Figure 2-8: **Bayesian optimization.** First row is the initialization with two points, and subsequent rows are the optimization steps. Left panels correspond to the GP while the right panels show the utility function with the guess for the next point indicated with black squares.

2.2.4 Shapley values

Many research questions are now tackled through data-driven machine learning methods due to their capacity to handle high-dimensional spaces. Unlike classical statistical methods that might generally be better suited for $n >> p$ situations and whose parameters can be interpreted, in most of machine learning methods it is often thought that the price to pay for increased prediction accuracy is the reduced interpretability due to the complexity of the models and their large number of parameters [132].

We can consider *interpretability* as the capability of describing the internal machinery of a model in such a way that is comprehensible by humans, whereas *explainability* refers to scrutinizing the reasons for which a model decides to make a specific prediction [133], commonly through the creation of post-hoc models or calculations that attempt to explain the first model. Although debate exists as to whether explainability is the correct paradigm to understand what a model is doing and even to whether the interpretability vs. accuracy tradeoff mentioned above is real (see, for example [134]), methods for ML explainability have been successfully used in bioinformatics applications. Examples include the characterization of gene expression in children in response to air pollution [135] and recovering features that are relevant for splice site prediction [136]. Here, we discuss a framework for ML model explainability based on Shapley values, which is a cooperative game theory concept introduced by Lloyd Shapley in 1951 (see [137] for a reprint of the original work).

Shapley values are a way to numerically summarize the value of playing a game in terms of expected payout for each player in a coalitional game. To formalize this, consider a subset of $S \subseteq F$ of players, where F is the set of all players, and a characteristic function $v : 2^N \rightarrow \mathbb{R}$ that maps a subset of N players S to a scalar value that determines the worth of the coalition $v(S)$ as the expected payoff sum obtained by the players in S. Shapley values are then the contribution $v(S \cup \{i\}) - v(S)$ of a player i to the coalition, averaged over all the possible arrangements for a coalition:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)}{|F|!}(v(S \cup \{i\}) - v(S)) \tag{2.39}$$

Shapley values were used in [138] to the context of deriving variable usefulness in multiple regression problems by comparing their contribution across all possible models. By adapting (2.39), we can compute the attribution for a feature i with a model $f_{S\cup\{i\}}$ trained including that feature and a model f_S without feature i:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)}{|F|!}(f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_s(x_s)) \tag{2.40}$$

with x_s being an observation with a subset of features S. Shapley values are described in [139] to be part of a class of methods for *additive feature attribution*: consider a model f for which explanations want to be derived, and g a model to explain the predictions of f; this class of methods then seeks to explain $f(x)$ on the basis of an observation x, and rely on a mapping function h_x that is particular to that observation to obtain a simplified version x' of x: $x = h_x(x')$. Then, g is constructed in such a way that $g(z') \approx f(h_x(z'))$ for $z' \approx x'$. The explanation model is the sum of attributions for informative features in this simplified input $z' \in \{0,1\}^M$, with $\phi_i \in \mathbb{R}$ as in (2.39), with h_x mapping the indicator to the original input space:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2.41}$$

Three desirable properties for additive feature attribution methods are covered in [139] together with the presentation of the following theorem for an explanatory model g that satisfies those properties as well as (2.41):

$$\phi_i(f,x) = \sum_{z'\subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z'\setminus i)] \tag{2.42}$$

This is the SHAP importance measure, which, in other words, corresponds to the Shapley values of a conditional expectation of the model: $f_x(z') = f(h_x(z')) = E[f(z)|z_s]$. One can see from (2.42) that this is prohibitive in terms of computation for large M, since all possible subsets need to be evaluated. Sampling workarounds have been proposed [140] (for Shapley values, although also applicable here), as well as kernel approximations that reduce the number of evaluations. Model-specific approx-

imations also have been derived, for example, for the case of trees [141]. The SHAP framework has been used to generate explanations, for example, for the prevention of hypoxaemia during surgery [142]. Here, we perform an experiment to illustrate how SHAP values can be used in the context of phenotype prediction with gene expression with the GTEx dataset. Two types of skin samples are collected: sun exposed (SE) from the lower leg, and not sun exposed from the suprapubic part of the body. From Fig. 1-13 it can be seen that skin samples (coded in two shades of blue) are clustered together when using their complete transcriptional profiles.

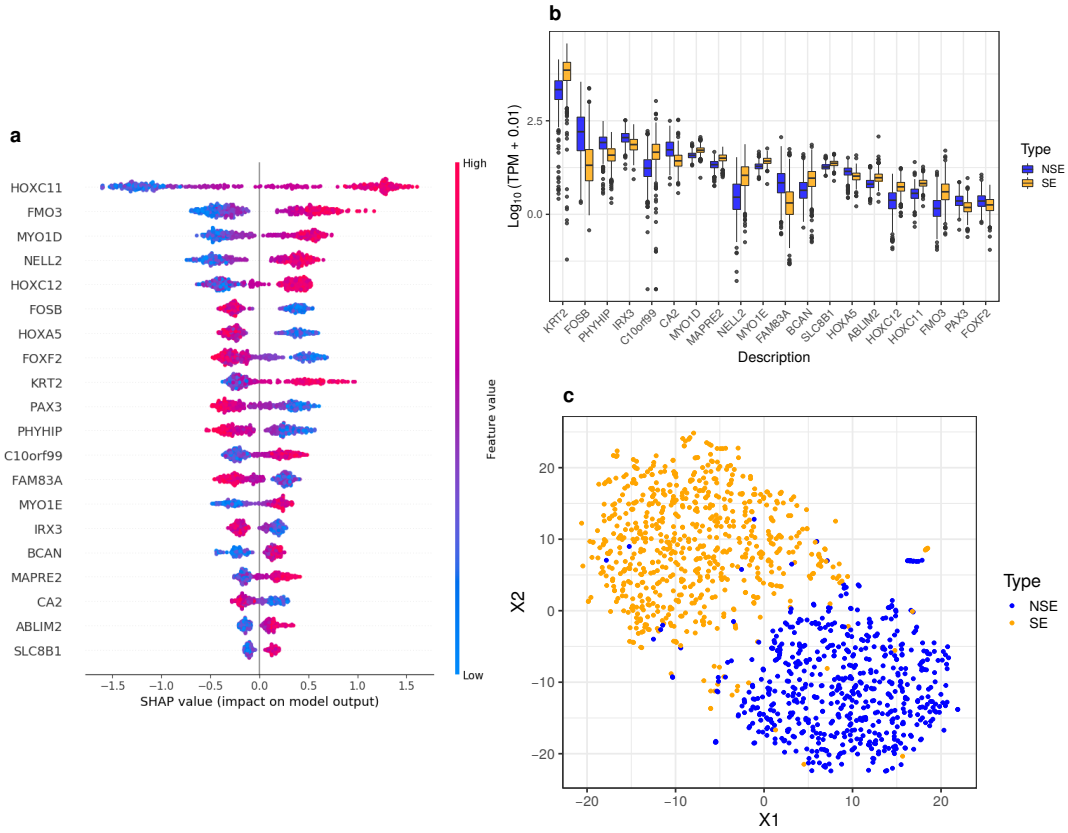


Figure 2-9: **SHAP values in skin type classification. (a)** SHAP values for the top 20 genes with more impact on model output. SHAP values are computed for every sample in every gene (each dot is a sample). Genes are sorted by the sum of absolute SHAP magnitudes across samples. Positive SHAP values contribute towards prediction of sun exposed skin, while negative values to the prediction of not sun exposed skin. **(b)** $\log_2(\text{TPM}+0.01)$ gene expression of the genes shown in (a), sorted by difference in their medians. **(c)** t-SNE of skin samples based on the expression of the 20 genes.

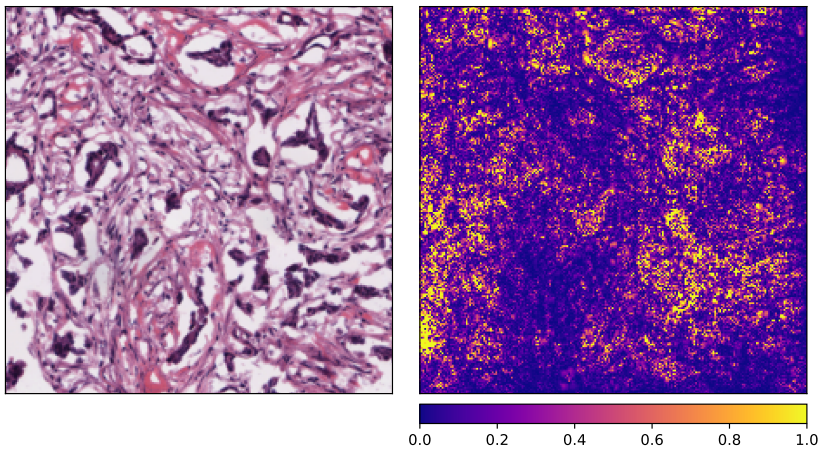


Figure 2-10: **Feature attributions for breast ductal carcinoma.** On the left, an histological image tile from a WSI of a TCGA breast sample affected by ductal carcinoma. On the right, the feature attributions generated through GradientSHAP.

We fit a cross-validated gradient boosted tree model to discriminate the type of skin sample (SE or NSE) based on the gene expression profile of each sample, and calculate SHAP values for each gene (Fig. 2-9a). The expression of these top-contributing genes is clearly different among skin types (Fig. 2-9b-c), demonstrating that, although similar at the global level, both skin types contain transcriptional signature subsets that allow to distinguish between types. In Chapter 5 we construct models to predict sex based on gene expression and analyze the sets of contributing genes based on SHAP values.

The SHAP framework can also be used to generate model explanations for non-tabular inputs, such as images, as well as to intepret model layers and neurons in neural networks. Here, we briefly illustrate such an usage: in Chapter 6 we fit a convolutional neural network to predict the tissue of origin from histological image patches derived from WSIs of cancer-affected tissues from The Cancer Genome Atlas [143]. By using GradientSHAP (gradient expectations by random sampling from a baseline), we can identify parts of the image that are highly predictive of the class, as shown in Fig. 2-10 where an image tile from breast tissue (affected by ductal carcinoma) appears on the left and its feature attributions appear on the right, highlighting desmoplastic stroma that is characteristic of this type of cancer.

2.3 Deep learning

The performance of statistical learning tools heavily depends on the set of features used to encode descriptions about our data. Although human-guided feature creation can certainly be of use in many cases, it becomes infeasible in tasks that deal with very high-dimensional data, such as images, text and temporal sequences, for which it is quite difficult to manually create meaningful features. For this reason, machine learning methods have been used not only to make use of feature sets to solve a particular problem, but also to learn an optimal data representation, process which is termed *representation learning*.

Deep learning (DL) has quickly found applications in many different fields, both in science and industry, due to its capabilities for representation learning, since it relies on representations structured through layers with varying levels of abstraction: Fig. 2-11 shows an example of a fully-connected feedforward neural network f (also called a multilayer perceptron) mapping an input $\mathbf{x} \in \mathbb{R}^{12}$ to an output $y \in \mathbb{R}^{1}$, with an input layer that corresponds to the number of features in the data, two intermediate (hidden) layers and a final output with a single output node, that could be used, for example, to model a regression problem.

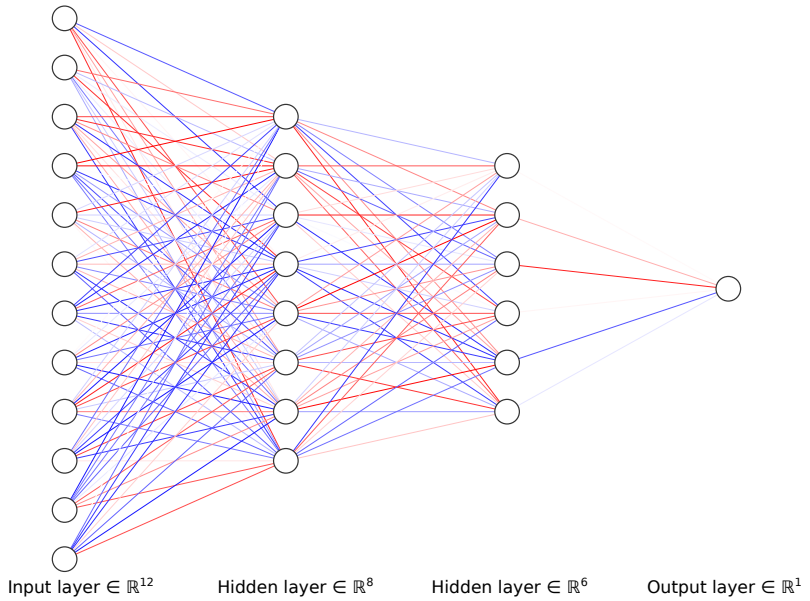


Figure 2-11: **Neural networks**. Fully-connected (FC) neural network with random weights, denoted by the edge color (positive = red, negative = blue).

This type of network is not considered DL, since it has a relatively shallow layout. The increasing depth (number of layers) of performant DL models is what originated the term deep learning, while the width of the network refers to the number of neurons in each layer. Nevertheless, feedforward neural networks serve to introduce the general concepts. As in all types of learning, they depend on a cost function as a proxy to approximate the underlying true model f^*. These networks make predictions based on composition of functions, for example, $y = f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$ with one function for each layer in the network shown before, and with the functions belonging to a class of transformations $\phi(\mathbf{x}, \boldsymbol{\theta})$ [144]. Here lies the main advantage of deep learning when compared to several classical ML methods: the mapping function ϕ performs a non-linear transformation of the input $\mathbf{x}$, allowing to learn complex intermediate representations from the input features at the cost of making this a non-convex optimization problem, with each layer having its own set of weights to parametrize the representation. In other words, ϕ roughly defines the architecture of the model, with each layer l having a set of parameters $\mathbf{W}^l$ and possibly biases $\mathbf{c}^l$ coupled with a layer activation function g_l that applies a non-linear transformation to the data coming from the previous layer in the network. For example, in Fig. 2-11, the transformations computed by the hidden layers for a single sample $\mathbf{x}$ could be:

$$\begin{aligned} \mathbf{h}^1 &= g_1(\mathbf{W}^1\mathbf{x} + \mathbf{c}^1) \\ \mathbf{h}^2 &= g_2(\mathbf{W}^2\mathbf{h}^1 + \mathbf{c}^2) \end{aligned} \tag{2.43}$$

With $\mathbf{W}^1 \in \mathbb{R}^{8\times12}$, $\mathbf{x} \in \mathbb{R}^{12}$, $\mathbf{c}^1 \in \mathbb{R}^8$ and $\mathbf{W}^2 \in \mathbb{R}^{6\times8}$, $\mathbf{h}^1 \in \mathbb{R}^8$, $\mathbf{c}^1 \in \mathbb{R}^6$. The choice of an appropriate activation function g_l depends on the problem domain [145] which is in turn related to the loss function used to evaluate the problem, the network architecture, and other design choices. Common activation functions are illustrated in Fig. 2-12. Examples of loss functions used in regression problems are L1-loss (mean absolute error), smooth L1-loss and mean squared error; in classification problems, the negative log-likehood loss, cross-entropy loss and Kullback-Leibler divergence. Other loss functions are used for problems such as ranking or learning embeddings.

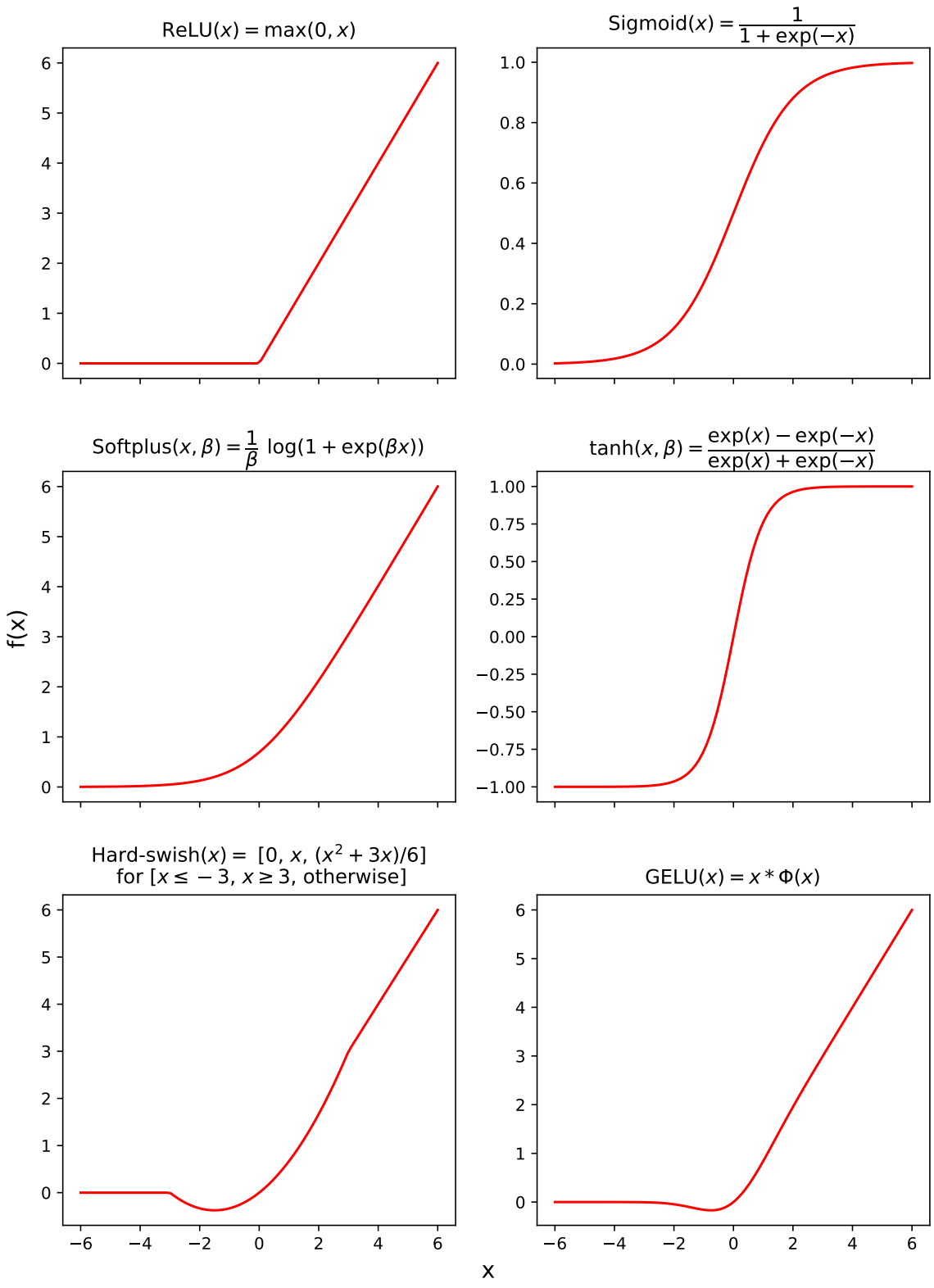


Figure 2-12: **Activation functions.** Common activation functions in deep learning applications. Φ is the CDF of the normal distribution. For a given input vector, these (particular) functions are applied elementwise.

Model training is performed in two stages: the *forward pass* in which the activations and pre-activations are computed for each layer, and *backpropagation*, which refers to the computation of the gradient of the layer weights and their optimization with a specific algorithm. Backpropagation was introduced in [146] and is the backbone of modern deep learning. Here, we illustrate the process following the exposition in [147], considering a quadratic loss function for a sample and target $(\mathbf{x}, y)$, with weights $\mathbf{W}^l$ for layer l and the same activation function g for all layers:

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}) = \left\| y - \mathbf{W}^l g(\mathbf{W}^{l-1} \dots \mathbf{W}^2 g(\mathbf{W}^1 \mathbf{x})) \right\|^2 \tag{2.44}$$

In the forward pass, the layer activations $\mathbf{h}_l$ are computed by applying g on the pre-activations $\mathbf{z}_l$ which are simply the linear combination of the layer weights and the activations from the previous layer, all the way until layer L:

$$\begin{aligned} \mathbf{h}^o = \mathbf{x}, \;\; \mathbf{z}^1 &= \mathbf{W}^1 \mathbf{h}^0 \\ \mathbf{h}^1 = g(\mathbf{z}^1), \;\; \mathbf{z}^2 &= \mathbf{W}^2 \mathbf{h}^1 \\ \vdots, \;\; &\vdots \\ \mathbf{h}^{L-1} = g(\mathbf{z}^{L-1}), \;\; \mathbf{z}^L &= \mathbf{W}^L \mathbf{h}^{L-1} \end{aligned} \tag{2.45}$$

In the backward pass, the gradient of the loss over the layer weights is computed. This depends a diagonal matrix $\mathbf{D}^L$ of derivatives of g evaluated at every pre-activation for each layer, as well as the derivative of the loss with respect to $\mathbf{z}^L$:

$$\mathbf{D}^l = \text{diag}(g'(\mathbf{z}^l_1), \dots, g'(\mathbf{z}^l_{d_l})) \tag{2.46}$$

$$e = \frac{\partial \mathcal{L}}{\partial \mathbf{z}^L} \tag{2.47}$$

With this, the gradients with respect to the weight layers are:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^l} = (\mathbf{W}^L \mathbf{D}^{L-1} \dots \mathbf{W}^{l+2} \mathbf{D}^{l+1} \mathbf{W}^{l+1} \mathbf{D}^l)^T e (\mathbf{h}^{l-1})^T, \;\; l = \{1, \dots, L\} \tag{2.48}$$

If we consider the following formulation for error backpropagation:

$$
\begin{aligned}
e^L &= e \\
e^{L-1} &= (\mathbf{D}^{L-1}\mathbf{W}^L)^T e^L \\
&\dots \\
e^1 &= (\mathbf{D}^1\mathbf{W}^2)^T e^2
\end{aligned}
\tag{2.49}
$$

then (2.48) can be simplified:

$$
\frac{\partial \mathcal{L}}{\partial \mathbf{W}^l} = e^l (h^{l-1})^T,\ l = \{1, \dots, L\}
\tag{2.50}
$$

Updates for each $\mathbf{W}^l$ are performed in the backward pass with stochastic gradient descent or related optimizers such as AdaGrad, Adam, and RMSProp. A lack of proper initialization of the model weights before starting the optimization can result in failure to reach convergence and training the network overall, and thus, several weight initialization schemes have been developed to prevent this [148]. In practical applications, data is usually large enough that is not possible to compute these for all the samples at the same time in a single pass. Thus, they are split in *mini-batches*, which refers to blocks of samples used to perform the forward and backward passes. The mini-batch size can be chosen, for example, with respect to hardware capabilities or aspects related to the optimizer, such as the learning rate [149]. The flexibility in formulating a model architecture is one of the reasons for the adoption of DL in many different types of computational biology problems, including: health record processing, gene expression signature extraction, sample cluster characterization, sample classification, inferring tissue-specific splicing mechanisms, TF-DNA motif recognition, promoter and enhancer identification, prediction of miRNA targets, secondary protein structure prediction, identification of protein-protein interaction networks, and lastly, bioimage analysis for the determination of morphological phenotypes [150], which is the primary subject of interest for this thesis. Since this strongly depends on the principles of CNNs, these will be briefly introduced in the next section.

2.3.1 Convolutional neural networks

CNNs, as opposed to vanilla feedforward networks, are meant to work with multi-way array inputs such as images, text, and video, among others, and are based on four properties of natural signals: local connections, shared weights, multiple layer usage, and pooling [151]. The pioneering concepts for Convolutional Neural Networks (CNNs) were introduced in the 80's, when LeCun et al. [152] explored the use of back-propagation in a network architecture to the problem of recognizing hand-written zip code digits, although some of the key components of CNNs like convolutional and downsampling layers were introduced earlier and are biologically inspired on the hierarchy model of the visual nervous system [153]. Later on, it was shown that an improved version of the original model, known as LeNet [154] (see Fig. 2-13) outperformed other models at that time, which kickstarted the field of deep learning.

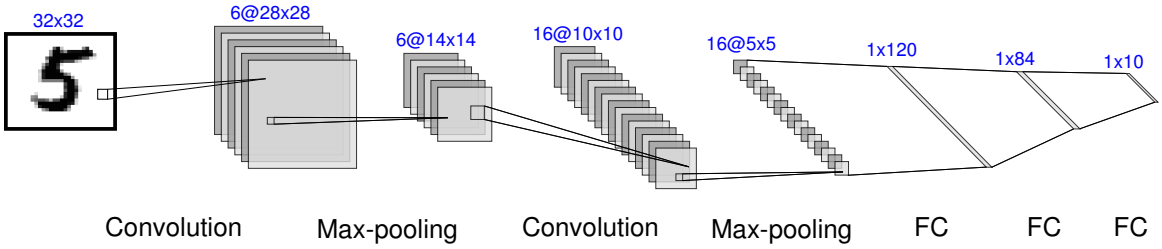


Figure 2-13: **Convolutional Neural Networks.** Example of a classical CNN: the LeNet architecture for handwritten character recognition. Input data is in the form of 32×32 (pixels) grayscale images. Through a combination of convolutions and max-pooling (subsampling), the network generates a final feature vector of size 10.

Convolutional layers are based on *feature maps*, which allow to learn specific characteristics of the input image (for example, in Fig. 2-13, after the first convolution, there are 6 feature maps of size 28x28), with all the neurons in a given feature map sharing the same parameters when scanning the input from the previous layer, but with each filter having a different set of weights, which makes each filter capture different information from the input. Pooling layers perform a smoothing of the input by calculating a summary statistic (such as the maximum value of a group of pixels), which can be thought of as a kind of downsampling.

Since LeNet, many major CNN models have been developed for classification, object detection, segmentation, among other applications, each one with a different architecture and hardware requirements. Examples of these architectures are: AlexNet, GoogleNet, VGGNet, ResNet and DenseNet [155]. Several techniques exist to improve the stability of the learning process. Examples are *batch normalization*, which refers to the normalization of feature maps in the network to have a mean of zero and unit variance; and *dropout* which is a type of regularization consisting in randomly removing network edges or neurons with a specific probability with the aim of reducing overfitting [156].

One might think that the applicability of deep learning is limited in situations when data is scarce to train a model. However, through the use of techniques like *transfer learning*, we can exploit the fact that the early layers of a model learn generic features (in the case of image data) such as edges and blobs [157], and thus, the weights of an existing trained network (for example, trained over ImageNet) can be transferred and adapted to suit a different problem or image types. This effectively reduces the sample size needed to train a model. Coupled with other techniques such as *data augmentation* which can be used to make transformations on a set of images (for example, random rotations, perspective warping, cropping, flipping, resizing, etc.), training a model to solve a domain-specific task does not necessarily have to depend on prohibitive amounts of data.

2.3.2 Integrating histopathology with molecular features

In Section 1.1.6, the different types of tasks that could be addressed with histological image data and deep learning models were introduced. Here, we will focus on describing recent advances that aim to produce general models of histology as well as the integration of histological image features with molecular features.

It is well known that histological features enable pathology characterization and classification into subtypes, although mixed evidence exists about how these features can be prognostic or predictive with respect to clinical outcomes in disease, for example, in non-small cell lung cancer [158] or melanoma [159]. However, recent work has used DL models trained on tile-based histological features to predict genetic features associated with disease. An example of this is in Kather et al. [160], where DL models trained on H&E histology are used to predict microstatellite instability (MSI). The latter is commonly identified through genetic analyses or immunohistochemistry, and aids in determining the response to immunotherapy in patients with gastrointestinal cancer. They find that the models can distinguish MSI-associated features, validating the findings in an external H&E set, and that the fraction of MSI-predicted tiles correlates both with gene expression and immunohistochemical data, suggesting that histology could play a role in inferring molecular features. Direct prediction of genomic alterations from histopathological patterns has also been addressed, for example, for whole-genome duplications and point mutations in driver genes [161]. Other works have focused in predicting gene expression patterns directly from WSIs and their spatialization in the context of cancer [162].

In the context of normal human tissues, existing work on an early version (v6) of the GTEx dataset has used convolutional autoencoders and sparse canonical correlation analysis to find image features relating to population variation and associated with genetic variants in stomach and colon tissues [163], suggesting that histological image features can be encoded and related with molecular traits, potentially to the level of spatially resolving gene expression as we present in early explorations in Chapter 6.

Besides inferring single traits from WSIs, these have also been used in conjunction with paired data types to improve predictions. For example, Chen et al. [164] developed a framework that integrates histopathology with genomic features such as mutations, copy-number variation and RNA-Seq for survival outcome prediction in cancer. To do this, they train models over each different data modality (CNNs and graph convolutional neural networks over WSIs, FCNs over genomic profiles) and combine the feature vectors through gating-based attention meachnisms and Kronecker products.

Existing models of histology in human are mostly focused on diseased tissues and identification of malignancies, but efforts have also been done to build models of normal tissue histology: Sing et al. [165] benchmarked the performance of different CNN architectures to recognize 46 different tissues from WSIs derived from rat tissue samples, finding that the generated feature vectors cluster together with respect to individual tissues and that morphologically similar tissues tend to have overlapping clusters. They suggest that these representations can be used to perform histological outlier identification, as well as discussing how these models, although not immediately applicable in their current state, could be used as a basis for cross-species histology predictions.

Most, if not all of these analyses, are built upon a set of common building blocks: WSIs are divided into a grid of smaller sized-tiles; models that are based purely on WSIs are trained used transfer learning (i.e. pre-trained architectures over existing datasets) or using (convolutional) autoencoders; feature vectors that encode the image patterns are generated as the output of the network and used for specific tasks. Although model design choices are variable in the literature (chosen WSI resolution, network architectures, regularization techniques, model training frameworks, source model for transfer learning, optimization, etc.), the biggest conceptual variation, at least in the case of linking histological features with molecular traits, lies in how the final feature vectors are used to perform the link. Some works decide to aggregate feature vectors from tiles corresponding to a single WSI into a single representation by calculating summary statistics for each component and then directly computing

sample correlations with other data pairs (but this complicates the traceability of the contribution of individual tiles), while others perform more complex meta-learning steps to directly model the link with the molecular features.

Although these types of tools have the potential to aid in disease screening with reduced costs, their application in clinical settings is not immediate and has challenges both from the ethical and legal perspectives. Gerke et al. [166] comment that there are several issues that need to be addressed to apply artificial intelligence models to healthcare: from the ethical standpoint, data consent, model safety and transparency, algorithmic fairness and biases; and from the legal side, concerns exist about safety and effectiveness, liability, among other issues.

Chapter 3

The effects of death and post-mortem cold ischemia on human tissue transcriptomes

Characterizing gene expression in tissues of living organisms is currently not possible due to the difficulties involved in obtaining these samples. For this reason, tissue samples obtained from post-mortem organisms are a proxy to study the behavior of gene expression, with the caveat that post-mortem RNA patterns cannot be considered exactly analogous to normal RNA levels in the living, since death and the post-mortem interval (PMI, which is the time elapsed since death) causes a cascade of events that alters these levels, and this degradation occurs in a tissue-specific manner. Here, we describe how post-mortem interval impacts gene expression levels and how these change across PMI intervals in 36 different tissues from 540 human donors. Our results suggest that transcriptional regulation continues after human death. Since the transcriptome responds to death in a tissue-specific manner, we used this variation to build models for estimating PMI for human individuals. We find that a few accessible tissues can be used to derive these estimates, and propose an usage protocol for an hypothetical real life scenario. Although larger sample sizes are needed to fully describe the predictive accuracy, these models suggest that gene expression signatures can carry information about the elapsed time since death.

Summary of my key contributions

- Built gradient boosted tree models to predict sample post-mortem interval from gene expression:

    1. Fit repeated cross-validated models to estimate PMI at the sample level, separately for each of 36 human tissues.
    2. Combined the tissue predictions to derive a prediction of the post-mortem interval at the level of the individual.
    3. Built separate models for blood samples to assess the possibility of overfitting, since these were sourced from ante- and post-mortem individuals, and reasoned that if there was a large amount of overfitting it would be possible to predict time to death. Found that ante-mortem predictions were essentially random.
    4. Performed a model stability analysis by evaluating sample prediction performance at the tissue level independently, through repeated cross validation to generate the distribution of model statistics.
    5. Found that a few readily-accessible tissues are enough to predict PMI, and proposed an example of a procotocol to follow to predict individual PMI in a real case scenario.
    6. Repeated the PMI prediction experiment using Transcript Integrity Numbers (TIN) and found similar prediction accuracy, although the sets of genes contributing to the predictions intersect only moderately, suggesting that effects of PMI on the transcriptome can result both from RNA degradation and regulation of gene expression.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-02772-x OPEN

The effects of death and post-mortem cold ischemia on human tissue transcriptomes

Pedro G. Ferreira *et al.*[#]

Post-mortem tissues samples are a key resource for investigating patterns of gene expression. However, the processes triggered by death and the post-mortem interval (PMI) can significantly alter physiologically normal RNA levels. We investigate the impact of PMI on gene expression using data from multiple tissues of post-mortem donors obtained from the GTEx project. We find that many genes change expression over relatively short PMIs in a tissue-specific manner, but this potentially confounding effect in a biological analysis can be minimized by taking into account appropriate covariates. By comparing ante- and post-mortem blood samples, we identify the cascade of transcriptional events triggered by death of the organism. These events do not appear to simply reflect stochastic variation resulting from mRNA degradation, but active and ongoing regulation of transcription. Finally, we develop a model to predict the time since death from the analysis of the transcriptome of a few readily accessible tissues.

Correspondence and requests for materials should be addressed to P.G.F. (email: pferreira@ipatimup.pt) or to R.Gó. (email: roderic.guigo@crg.cat). [#]A full list of authors and their affliations appears at the end of the paper.

Post-mortem human tissue samples are a valuable resource for biological research. Specifically, use of post-mortem material is crucial for studying the patterns of normal gene expression underlying tissue specificity within individuals, as sampling such tissues from living individuals would be impossible. However, the death of an organism triggers a cascade of events that ultimately, in a relatively short time frame, lead to cell death and autolysis. Although DNA is known to be relatively stable over long post-mortem periods, RNA is much more labile in nature, and sensitive to degradation in a tissue-specific manner[1]. There are conflicting reports on how the post-mortem interval affects RNA integrity[2–10] but several studies, in different mammals, have shown that RNA can remain largely intact even for considerable time periods, when samples remain properly stored. In addition, a variety of pre-mortem factors, including environmental parameters and the circumstances of death, may also influence the quality of the collected tissues and their RNA[8,11]. RNA quality impacts measures of gene expression. Recent studies[12–15] have shown that sequencing lower RNA quality samples, as measured by the RNA integrity index (RIN)[16], leads to a decrease in the quality of the data obtained by high throughput RNA sequencing (RNA-seq), and the use of RIN, and other related variables, as covariates in differential expression analysis, has been recommended[12,13,17].

On the other hand, transcriptional changes are expected to occur as a response to the death of an organism. However, little is currently known about how death and the length of the post-mortem cold ischemia interval specifically affect gene expression since most existing reports are based on very few genes, tissues or individuals[5–7,10,11,17,18]. Therefore, RNA levels measured in post-mortem tissue samples will be affected both by biological responses to organism death, as well as to RNA degradation occurring as a consequence of cell death. Understanding how these effects are dependent on the post-mortem interval is essential for the proper use of post-mortem gene expression measures as a proxy for ante-mortem physiological gene expression levels[5,10,18–20].

Here we analyze the GTEx[21–25] RNA-sequencing data to investigate the impact of death and the post-mortem cold ischemic interval on the transcriptomes of human tissues. We find that different tissues have a different response over the time elapsed since death, but that when appropriate covariates are identified and taken into account, the impact of death on tissue transcriptomes can largely be controlled. We identify the cascade of molecular events triggered by death specifically in the Blood transcriptome. Finally, we develop a model to predict the time since death from the analysis of the transcriptome of a few readily accessible tissues.

Results

Study overview. We used mRNA sequencing data from the GTEx project (V6, Supplementary Table 1 and 2), and the derived gene and transcript quantifications obtained on Gencode[26] V19. We restricted our analyses to 36 tissues with >20 samples, including whole blood and two brain sub-regions (cortex and cerebellum) for a total of 7105 samples, corresponding to 540 donors (Supplementary Fig. 1, 2, 3, Methods). All samples were collected and preserved with the PAXgene Tissue preservation system[21].

The GTEx metadata contains an extensive annotation of samples and donors, including the postmortem interval (PMI). For GTEx individuals, PMI is defined as the time since death to the start of the GTEx collection procedure. For tissue samples, this is defined as the time in minutes spanning the window from the moment of death, or the cessation of blood flow, until tissue stabilization and/or preservation takes place, with values ranging from 17 to 1739 min (Fig. 1a, Supplementary Note 1). Correlation analysis shows that there is a strong association of PMI with variables describing tissue recovery and death circumstances, as these variables are correlated and reflect the same intrinsic features of the collection procedures (Supplementary Fig. 4, Supplementary Table 3). The relationship between PMI and RNA stability is very tissue-dependent (Fig. 1b, Supplementary Fig. 5, Supplementary Table 4), in agreement with previous observations[5,17,27].

Impact of PMI on gene expression. To identify genes that changed expression depending on PMI, we used the five PMI intervals also used by the GTEx Biospecimen Methodological Study (BMS)[21], and asked which genes had a significant and noticeable change between two consecutive time intervals (>2-fold change and Wilcoxon test $p < 0.05$, see Methods and Supplementary Note 2). The number of genes with a significant change in at least one interval transition varies widely between

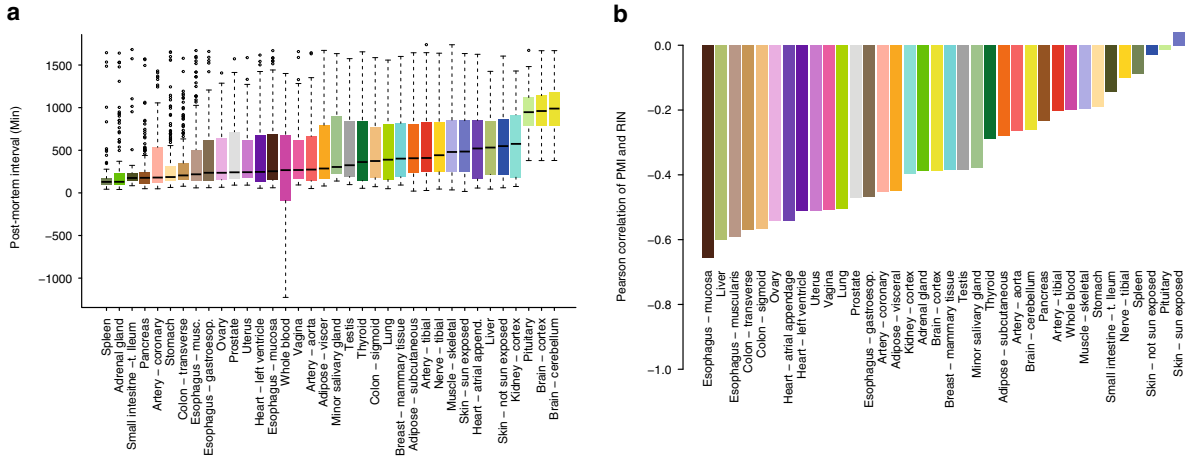


Fig. 1 Characteristics of the samples and tissues used in this study. **a** Distribution of PMI values (in minutes) with tissues ordered by the median value. Whole blood contains samples with negative time corresponding to samples obtained pre-mortem. **b** Distribution of Pearson correlation between PMI and RIN values. Esophagus, Liver, Colon, Ovary, Uterus, Vagina, and Heart are the tissues in which RIN is more affected by PMI ($r < -0.5$), while Skin, Pituitary, Spleen and Nerve are the ones in which is less affected ($r > -0.1$)

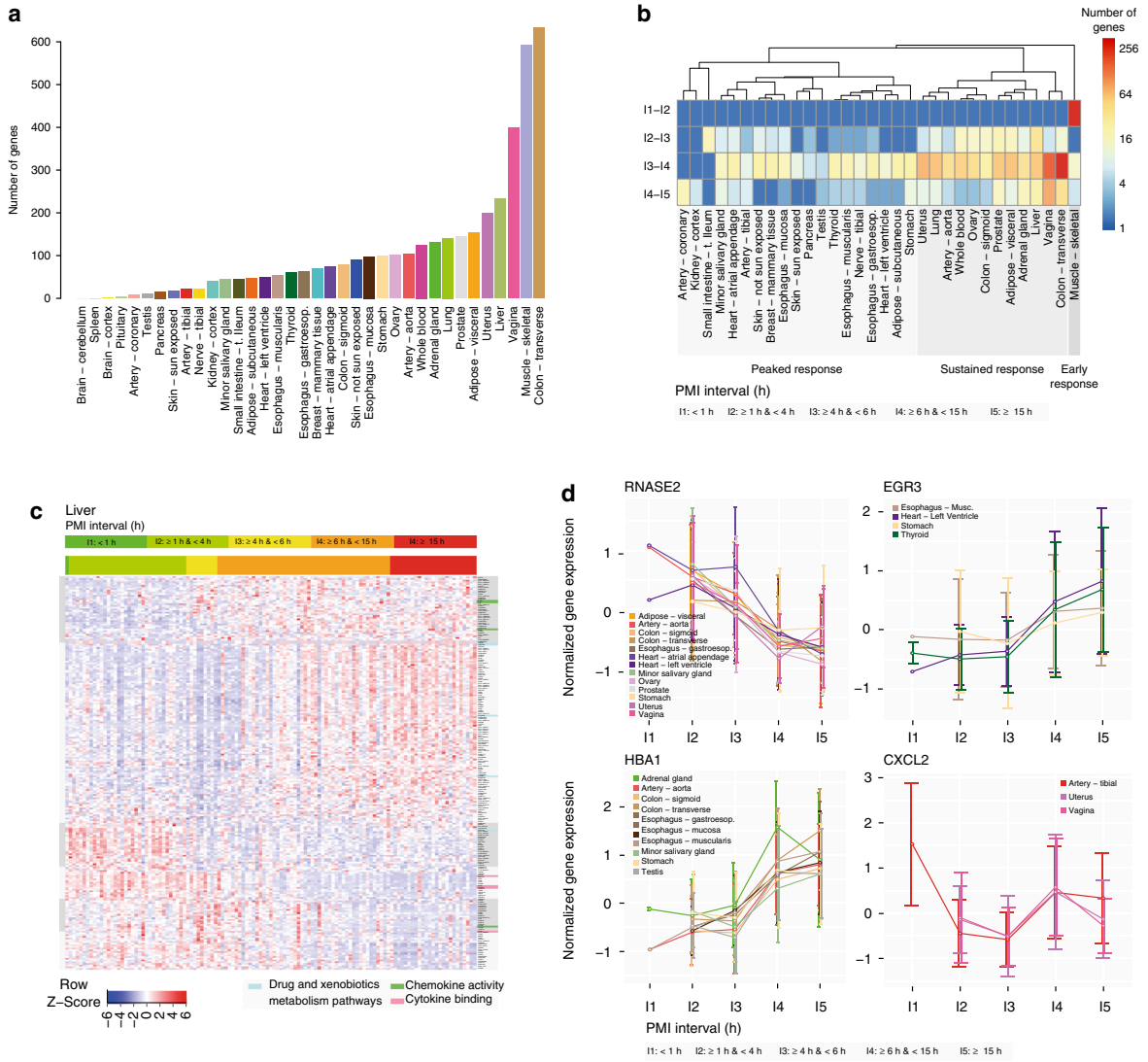


Fig. 2 Effect of PMI on gene expression. **a** Distribution of the number of genes with significant temporal changes per tissue between at least two time intervals. Brain and pituitary have longer PMIs, only within last two time intervals, thus less interval ranges to detect significant changes. **b** Heatmap with the number of genes with significant changes detected between two consecutive time intervals. **c** Heatmap with normalized expression values for genes with changes in liver. The top bar is the color code for the PMI interval of each sample, the right bar list genes involved in the various functions and pathways. On the side we highlight sub-clusters with different patterns of temporal expression. **d** Example of four genes with different temporal patterns. *RNASE2* is a non-secretory ribonuclease involved in several functions; *HBA1* is alpha hemoglobin involved in oxygen transport; *EGR3* is a transcriptional regulator involved in early growth response; *CXCL2* is a chemokine gene that encodes secreted proteins involved in immune and inflammatory processes

tissues, ranging from none in brain cerebellum and spleen, to >600 in muscle and colon transverse (Fig. 2a). Although most tissues are characterized by a sharp shift in gene expression at around 6 h after death, there are remarkable differences between tissues regarding the transcriptional response to PMI (Fig. 2b). Some tissues (e.g., muscle) exhibit an early response, with most genes that change expression doing so right after death (Supplementary Fig. 6). Another set of tissues show a more sustained response, with gene expression changes of similar magnitude occurring through all PMI intervals (Fig. 2c). Finally, another set of tissues show a peaked response, with most changes occurring between the intervals of 4–6 h and 6–15 h (Supplementary Fig. 7).

There is little overlap of affected genes across the tissues. We identified 187 genes (94 are protein-coding) with post-mortem gene expression changes in at least three tissues (Supplementary Fig. 8). The gene that showed consistent changes across the largest number of tissues was *RNASE2*, a gene from the family of ribonucleases, enzymes involved in the degradation of RNA. *RNASE2* shows a consistent decrease in expression across 13 tissues (Fig. 2d). Two alpha globin genes, *HBA1* and *HBA2*, involved in the transport of oxygen from the lung to the peripheral tissues, show an increased expression in several tissues but not in blood, where they are the most expressed genes (Fig. 2d). Several histone genes show increased patterns of expression in line with previous results[28,29] (Supplementary Fig. 8, Supplementary Data 1). Growth factors, such as *EGR3* also have an increased expression from 4hr to later on (Fig. 2d, Supplementary Fig. 8). Other genes such as the chemokine

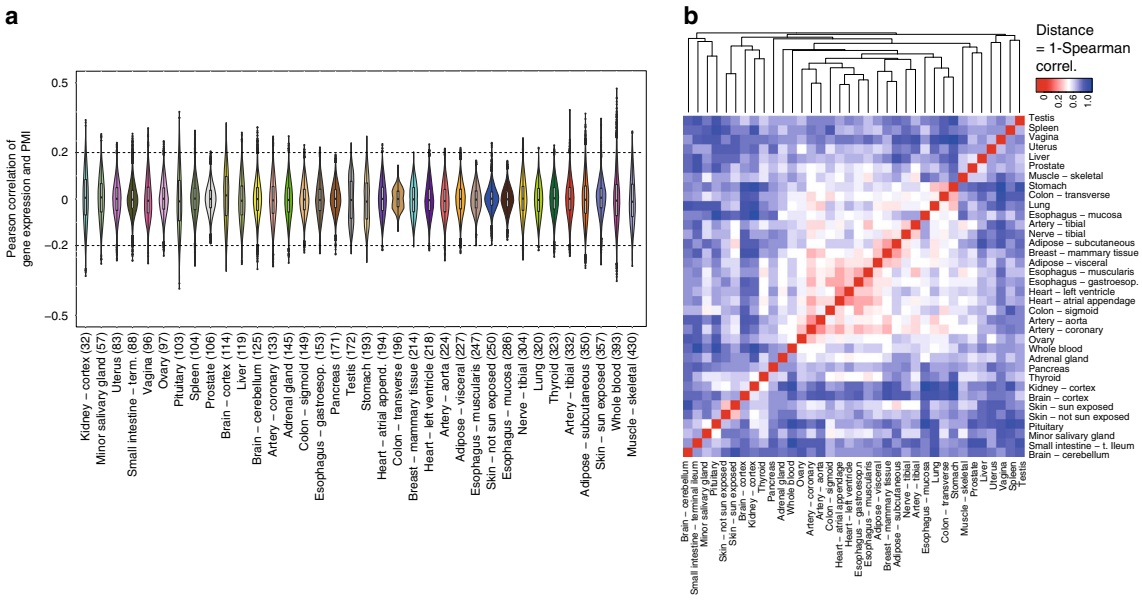


Fig. 3 PMI and gene expression correlation patterns. **a** Distribution of Pearson correlation between gene expression and PMI, across the different tissues (sorted by sample size, in parenthesis). Only for a few genes, this correlation exceeds an absolute *r*-value of 0.2. **b** Clustering based on the ranking (Spearman) correlation of the values in (**a**) show that sub-tissues of a given tissue or closely related organs have the similar patterns of correlation

CXCL2 show a more dynamic behavior with expression changes in opposite directions at subsequent intervals (Fig. 2d). Gene ontology analysis of the genes affected across several tissues (Supplementary Fig. 8) shows enrichment for genes in the extracellular region and genes involved in nucleosome and chromatin assembly and in protein–DNA complexes. There is also enrichment for inflammatory and immune response processes.

While there are noticeable changes in gene expression associated with PMI, we nonetheless found that the characteristic transcriptional signature of tissues remains largely intact through the PMI intervals considered here. We clustered the GTEx samples at these intervals and measured, using modularity (see Methods), how well the clustering recapitulates tissue type. Here, we compute modularity on the network constructed from gene expression correlations between samples when the data are grouped by tissues. Modularity remained stable through the PMI intervals at any threshold of the correlation defining the network edges (see Supplementary Fig. 9).

Because PMI dependent expression changes are largely tissue-specific, they could confound tissue differential gene expression since the observed effects could be caused by differential response to PMI rather than by differences in tissue biology. To investigate to what extent these effects can be controlled for, we used a linear regression model that allows incorporating additional covariates. We specifically selected fourteen variables, predominantly demographic, medical history and sample QC metrics that are orthogonal to the sample collection procedure, to include as expression covariates in the model[21] (see Supplementary Fig. 4, Supplementary Table 5 and 2). These are essentially the covariates employed in the GTEx eQTL analyses[24]. Residuals were then used as the expression phenotype and the Pearson correlation (r) as a measure of linear relationship with PMI (Methods, Supplementary Notes 2). On average we found only 54 genes per tissue (0.2%), which showed significant correlation of gene expression with PMI (FDR < 1%) (Fig. 3a, Supplementary Table 6), compared to 6919 genes per tissue (39.3%), if using the same model without covariates. In most of these cases, however, the effect is small (only 189 (1.1%) with $r < |0.2|$ (Supplementary Fig. 10). Moreover, clustering of tissues based on the ranking of correlations gene expression-PMI generally recapitulates tissue type (Fig. 3b). These results suggest that the effect of PMI on measured gene expression is relatively modest and can be further minimized by using appropriate covariate correction in analyses. The effect is weakly mediated by the number of exons, the length of the gene and of the coding region and GC content (Supplementary Table 7). PMI has also little effect on the proportion of intergenic RNA-seq reads, as well as on 3' mapping bias, commonly observed in RNA degraded samples[12,14,30–32] (Supplementary Figs. 11–15, Methods).

To specifically analyze the impact of PMI in energy metabolism, we investigated its relationship with mitochondrial RNA (mtRNA) levels. We observed no significant changes in mtRNA concentration across different RIN values, and donor ages (Supplementary Figs. 16 and 17). Across most tissues, samples exhibit a significantly lower proportion of mitochondrial reads in late PMIs (Fig. 4a, Methods), except blood, salivary gland, heart-left ventricle and, particularly, liver that exhibits a substantial higher proportion of mitochondrial RNAs for late PMIs (Fig. 4b, Supplementary Fig. 17). Decreasing mtRNA abundance across all PMI intervals is observed specifically in female tissues (ovary, vagina, and uterus, see Fig. 4c).

Finally, we investigated the effect of PMI on splicing. We calculated the inclusion levels[33] of internal exons (Supplementary Fig. 18a, Methods). We then performed linear regression analysis of PMI and PSI values and found 1,399 exons (612 unique) significantly correlated with PMI ($|r| > 0.5$ and FDR ≤1%; Fig. 5a–c), of which 160 were observed in three or more tissues (Fig. 5a). In contrast to gene expression, there is a substantial sharing of exons among the top affected tissues (those with ≥ 20 significant exons), with the tissue pairwise overlap ranging from 43% to 82%, representing 22 to 76 shared exons. Functional analysis of genes with recurrent exons (i.e., with association with PMI in more than two tissues) shows a noteworthy enrichment

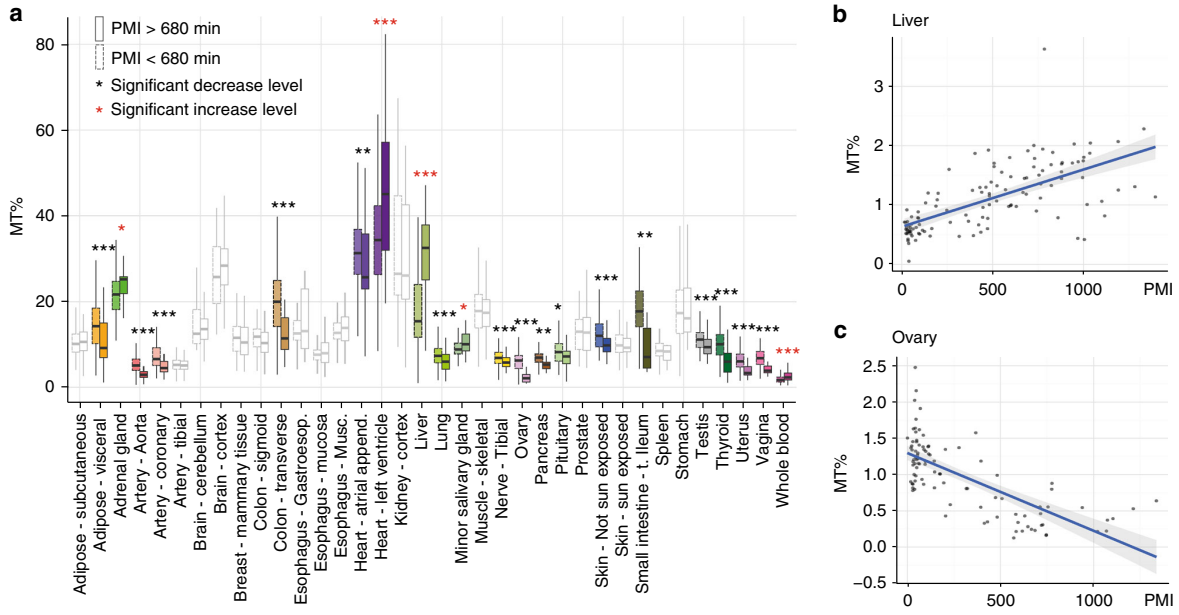


Fig. 4 Effect of PMI on mitochondrial transcription and splicing. **a** Proportion of RNA-seq reads originating from mitochondrial genes (mtRNA concentration) in early (≤680) and late (>680) PMI intervals. **b**, **c** mtRNA concentration depending on PMI in Liver (**b**) and Ovary (**c**)

on RNA binding and RNA splicing genes (Supplementary Fig. 18b, c). We also investigated if, as a consequence of death, we could observe a generic alteration of splicing. As a proxy for splicing alteration, we computed the Shannon's entropy on the relative abundance of a gene's alternative splicing isoforms (Methods)—higher values corresponding to more stochastic production of alternative isoforms. We did observe an increase of splicing entropy in many cases (Fig. 5d, e), although not a systematic trend across all tissues (Supplementary Fig. 19).

Changes induced by death in the whole blood transcriptome. Among the samples collected for GTEx, the blood samples are unique in having been collected pre-mortem for some donors and post-mortem for others. This provides an opportunity to assess the impact of death on the gene expression of a specific tissue. Dimensionality reduction (MDS) and hierarchical clustering of gene expression profiles clearly distinguishes pre- and post-mortem states of blood samples (Fig. 6a and Supplementary Fig. 20). The "cause of death" (assessed by the 4-point Hardy scale classification, Supplementary Notes 1) is quite different for individuals from whom Blood was obtained pre-mortem compared to post-mortem, but this does not appear to have a major impact on the clustering, which is independent of Hardy classification (see Methods, Fig. 6a).

To characterize changes in gene expression that are triggered by death, we identified genes that were differentially expressed between pre-mortem and post-mortem blood samples, the latter being collected at several different PMI intervals (Fig. 6b, Supplementary Table 8 and Supplementary Data 2). Immediately following death (and up to seven consecutive hours) we observe an increase in the expression of many genes, and a decrease in the expression of a few. The majority of the changes in gene expression, however, occur between 7 and 14 h post-death, with thousands of genes showing differential expression (equally in both directions) relative to pre-mortem samples. Then, between 14 and 24 h, the transcriptome seems to stabilize, with comparatively few genes showing differential expression relative to pre-mortem samples (among those that do, there are more over-expressed than under-expressed). Categorizing the nature of these changes in gene expression in blood samples following death, we observed five main functional activities[34] (Fig. 6c, Supplementary Table 9, Supplementary Notes 3): 1) changes in DNA synthesis and fibrinolysis; 2) deactivation of the immune response; 3) an increase in activity of processes related to cell necrosis; 4) an abrupt inactivation of carbohydrate metabolism, synthesis of lipids (e.g., cholesterol) and ion transport; and 5) an activation of processes related to Blood coagulation and Response to stress (Supplementary Fig. 21a). Specifically, the way in which carbohydrate metabolism is affected, with severe deactivation of the tricarboxylic acid cycle, while glycolysis is activated (FDR < 10^{-27}; Fig. 6d, Supplementary Table 9), suggests that hypoxia is likely playing a major role in the initial pre- to post-mortem transition (FDR 7.2×10^{-67}). More gradually, the immune system is also deactivated (several immunity-related functions with FDR < 10^{-30}, Supplementary Table 9, Supplementary Fig. 21b). In addition, a response to stress, along with the detection of DNA damage and the activation of the corresponding repair machinery is observed (FDR < 10^{-14}; Supplementary Table 9). Finally, a general arrest of cell proliferative processes occurs. Processes like growth arrest are activated and others, like Initiation factor, the starting process of protein production, are dramatically deactivated.

The transcriptional changes detected above may partially be related to changes in the cellular composition of blood triggered by death. Indeed, blood is a complex tissue composed of multiple cell types. We investigated differences in cell composition between blood samples collected pre- and post-mortem. We used CIBERSORT[35], to deconvolute bulk gene expression into expression levels for 18 different cell types. We found significant differences in overall cellular composition between pre- and post-mortem blood samples ($p < 0.001$), the most notable changes induced by death being an increase in resting NK cells and CD8 T-cells, and a substantial reduction in neutrophils (Fig. 7a). These results are consistent with the observed deactivation of the immune system (Supplementary Fig. 21b), since similar trends are observed to be associated with dysregulation of the immune

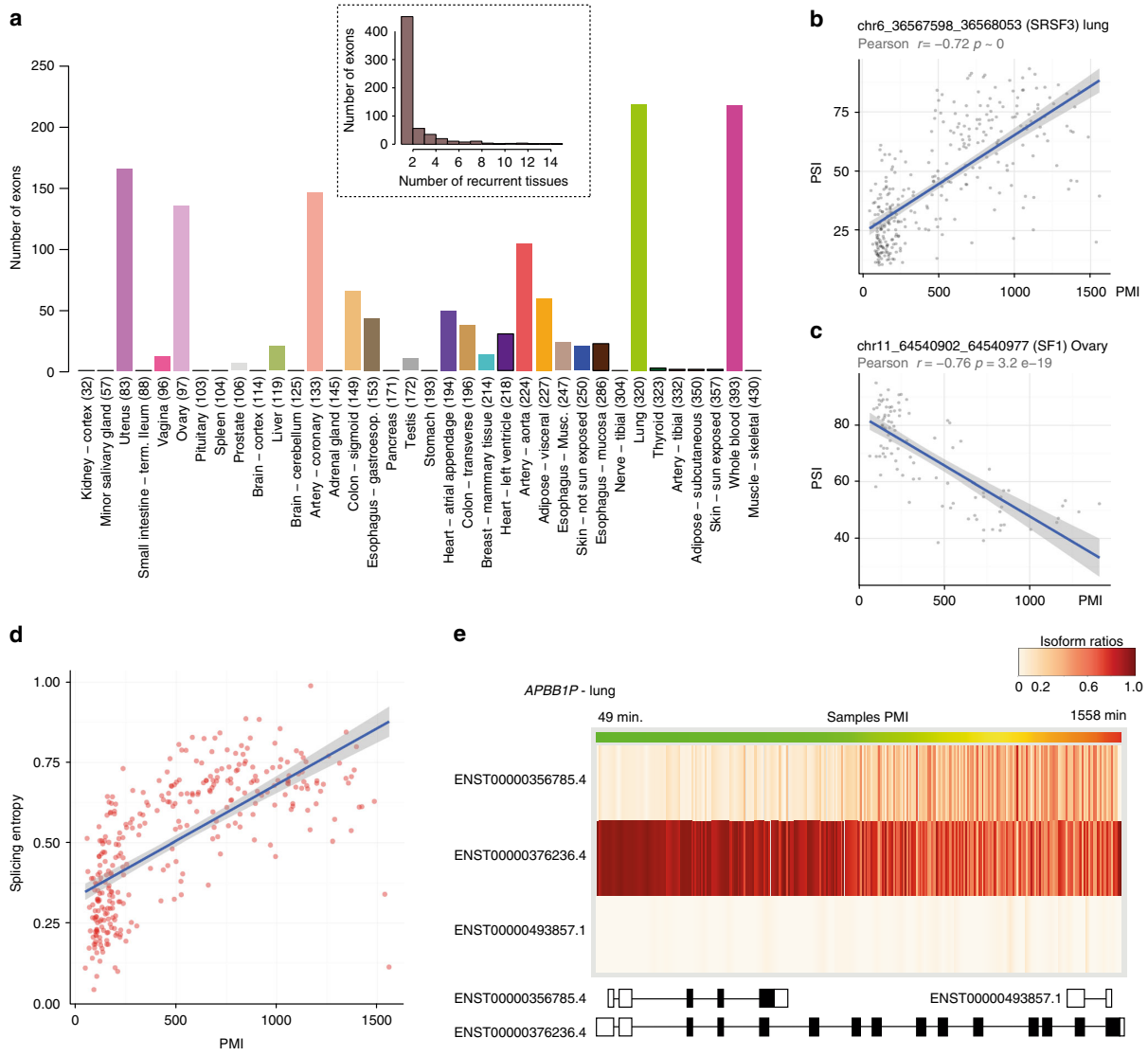


Fig. 5 Effect of PMI on splicing. **a** Distribution of the number of exons with significant differential inclusion across the different tissues; inset: number of exons with differential inclusion occurring in multiple tissues. **b**, **c** Examples of two exons with PMI-correlated inclusion levels. *SRSF3* and *SF1* encode pre-mRNA splicing factors that form part of the spliceosome. **d** Splicing Entropy for the *APBB1IP* gene depending PMI in the lung tissue. **e** *APBB1IP* gene has three isoforms. The heatmap shows the proportion of each isoform in the Lung samples sorted by PMI (1 indicates that only one isoform contributes to the expression of the gene and 0 that the isoform is not expressed) PMI of the samples range from 49 to 1558 minand are represented in the green to red gradient scale bar. Exon structure (not to scale) for the three isoforms is represented below. This figure depicts how the expression of the longer transcript in this gene becomes less dominant as PMI increases, and, as a consequence, the abundance of the different isoforms tends to converge

system with age. Neutrophils, in particular, are the first cells to migrate to pathogenic infected sites, and a decrease in their levels implies impaired ability to traffic into and out sites of infection.

Death also has an observable impact on splicing in the blood transcriptome. We identified 497 exons (from 381 genes) that were differentially included between the pre- and post-mortem samples ($p < 0.01$, $|\Delta PSI| > 0.1$, Supplementary Fig. 22, Supplementary Table 10). This represents 14% of all exons (3441) that were found to be variable across samples (Methods). Most of these 497 exons (75%) tended to be "included" in the pre-mortem samples (and not in the post-mortem samples), suggesting that splicing deregulation is occurring. Indeed we found that post-mortem samples have a higher entropy than pre-mortem samples (Fig. 7b), reflecting tighter, more controlled, usage of splicing isoforms in the pre-mortem samples. In general, we found there was an increased usage of the major (dominant) isoform in the pre-mortem samples relative to the post-mortem samples (Supplementary Fig. 23).

Prediction of the post-mortem interval from gene expression. The precise estimation of PMI is a problem of central importance in forensic pathology. Traditional methods for this task rely on physical modifications observed on the body, including algor, livor, and rigor mortis[36]. However, these approaches may be unreliable or inaccurate[18]. The use of RNA assays as an addition to the forensic tool kit is of growing interest with studies looking

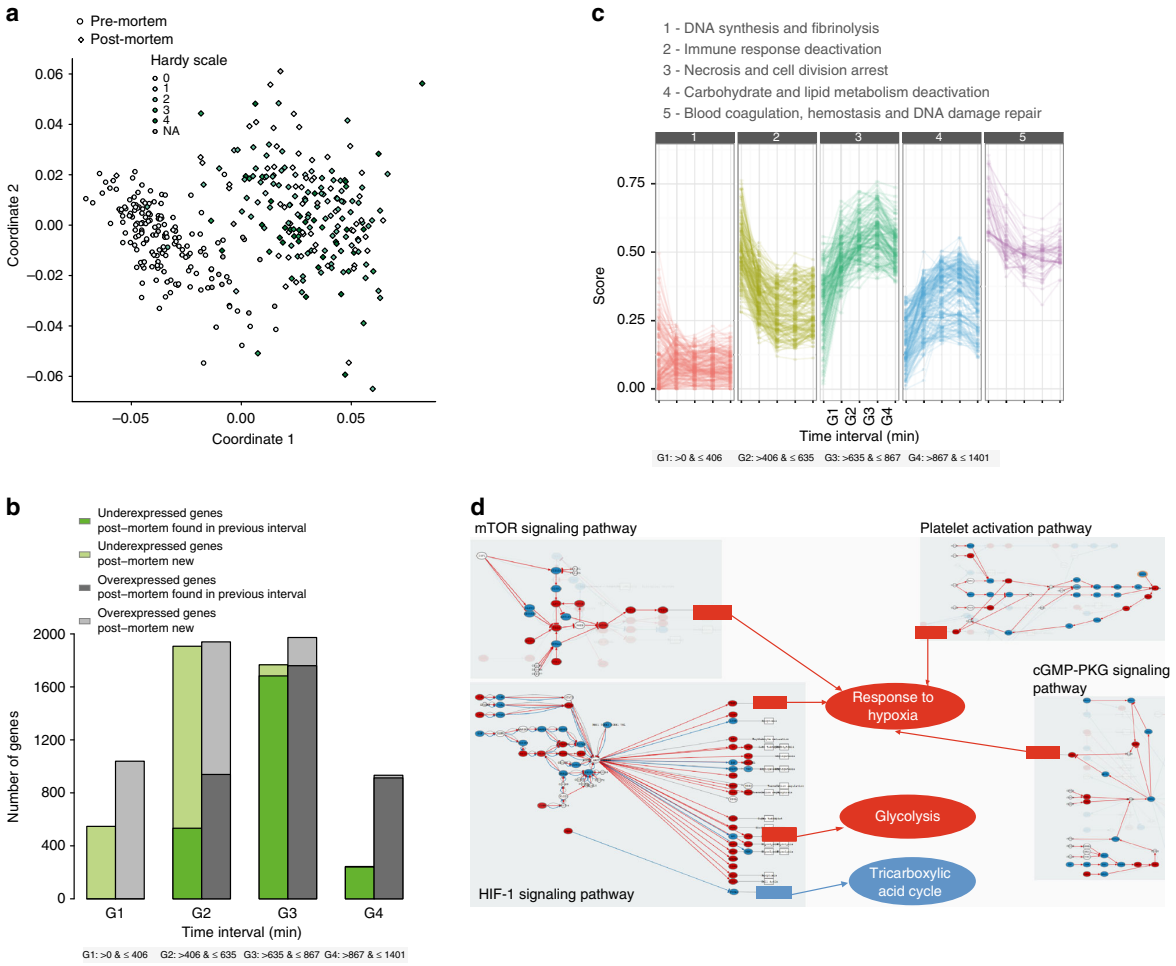


Fig. 6 Transcriptional changes in blood after death. **a** Multi-Dimensional Scaling of blood samples shows separation between pre and post-mortem samples. Samples are colored by the Hardy scale of the cause of death. **b** Number of genes differentially expressed between the pre-mortem samples and the post-mortem samples stratified at different PMI intervals. Darker filling corresponds to genes that are found as differentially expressed in the previous interval. **c** The five main temporal patterns of change in functional activities upon organismic death. **d** Hypoxia seems to play a major role in the pre-to post-mortem transcription as reflected in the way in which the carbohydrate metabolism is affected (activations in red, deactivations in blue). Response to hipoxia is activated from pathways "Platelet activation pathway" and "cGMP-PKG signaling pathway" through the activation of the corresponding circuits that end in the effector gene ITPR1, annotated as Response to hypoxia, and from pathways "HIF-1 signaling pathway" and "cGMP-PKG signaling pathway" through the activation of the effector gene VEGFA. The "HIF-1 signaling pathway" also activates Glycolysis through the activation of different circuits that trigger effector proteins (PDK1, PFKL, ALDOA, etc.) with annotations such as glycolytic process, canonical glycolysis, glucose metabolic process, etc. The "HIF-1 signaling pathway" also inhibits Tricarboxylic acid cycle through the inhibition of circuits that trigger the effector protein PDHA1 with diverse GO annotations such as tricarboxylic acid cycle, acetyl-CoA biosynthetic process from pyruvate or carbohydrate metabolic process

for a correlation between RNA degradation and PMI[1,10]. The use of mRNA markers in PMI prediction also holds great promise, but so far only a few genes from a handful of human tissues have been tested[18]. Herein, our analyses suggest that the patterns of gene expression change with time after death in a tissue specific manner, and might thus be collectively used to predict the PMI for a given individual. We use the GTEx RNA-seq data to develop and to test such an approach. We first use gradient boosted trees[37] to infer models that use expression of protein coding genes to predict the PMI of each tissue separately. We used data from 399 individuals (about 75% of the 528 available individuals) for training the models and 129 (~25%) for testing (Supplementary Fig. 24 and 25). In the test set we obtained R^2 values between predicted and real tissue PMI ranging from 0.78 to 0.16 (Supplementary Fig. 26 and 27), similarly to what was obtained in the training set (Supplementary Fig. 25). We also calculated for each sample, the difference between real and predicted tissue PMI, and found little deviation on average, although the models tend to overestimate PMI (Fig. 8a). To assess the possibility of overfitting due to the complexity of the data we performed a model stability analysis via resampling (Supplementary Fig. 28). In addition, we used the blood samples and we repeated the training and testing procedures (×100) separately in post-mortem and in pre-mortem samples. We reasoned that if predictions resulted from overfitting, we should be able to predict the time to death in pre-mortem samples equally well as the time since death in the post-mortem samples. Reassuringly, predictions of time to death were essentially random (median R^2 0.02 compared to 0.47 for predictions of time since death, see Supplementary Fig. 29).

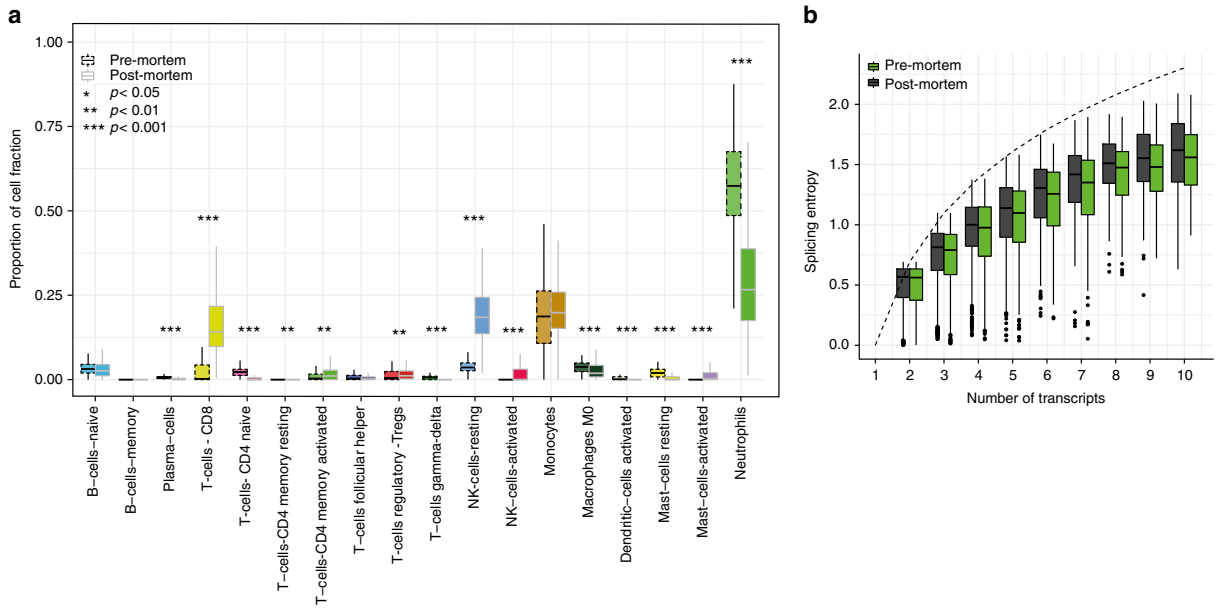


Fig. 7 Differential cellular composition and splicing entropy in blood. **a** Cellular composition analysis for 18 cell types shows an increase in NK-cells-resting and T-cells-CD8 and decrease in Neutrophils composition from pre- to post-mortem blood samples. **b** Splicing entropy in pre- and post-mortem samples for genes of different number of isoforms

To infer the PMI of each individual, we subtracted from each tissue predicted PMI the time elapsed since the beginning of the GTEx procedure to the processing of that tissue, and averaged the resulting values (Fig. 8c, d). On average, the PMI prediction error (signed difference between real and predicted) is 9.45 min and the median of −63.75 min (Supplementary Fig. 30). The R^2 of the predicted and real PMI is 0.77 when all tissues are considered, and 0.8 when using only the top 20 tissues ($R^2 > 0.5$ in the training set, Fig. 8b). As a measure of stability of the PMI prediction on a given individual, we assess the consistency of the tissue PMIs for the individuals. We reason that if all tissues predict consistently very similar PMIs, the prediction of the PMI for the individual is more reliable than if the tissue PMI predictions are very variable across tissues. To assess the consistency of the tissue PMI prediction for a given individual we compute the coefficient of variation (cv), lower values thus indicating more reliable predictions. Supplementary Figure 31 shows the cv distribution on the individuals from the test set.

Since the availability of so many tissues is unrealistic in a forensics scenario, we identified the smallest combination of tissues that can be used to determine an individual's PMI accurately. For each individual in the initial test set, we identified the subset of tissues of a fixed size that can predict the individual PMI with the highest precision. We find that for subsets of sizes 2–6, the tissues that appear more frequently are Adipose—Subcutaneous, Lung, Thyroid, and Skin (Sun Exposed) (Supplementary Fig. 32). We prioritized this approach over simply identifying the combination of tissues with highest R^2. Predictions using these four tissues are even superior to those using all top 20 tissues ($R^2 = 0.86$) (Supplementary Fig. 33), and actually only marginally superior to those obtained using some combinations of only two tissues among the four above (Supplementary Fig. 33 and 34).

We investigated to what extent the PMI predictions are robust to the causes of death since this could also have an impact on the transcriptome. To have sample sizes large enough, we grouped the causes of death reported by GTEx (Supplementary Fig. 35a) in three major death classes: cerebrovascular disease, heart disease, and other causes of death. We did not observe an impact of the class of death in the accuracy of the predictions, as measured by R^2 (Supplementary Fig. 35b).

The results above suggest that gene expression values (estimated, for instance, through RNA-Seq) can be used to effectively predict time since death. Figure 9 summarizes the main steps to follow in a putative real case scenario.

We also investigated whether estimates of RNA degradation can be used to predict PMI. We have employed exactly the same methodology, but using the transcript integrity number[15] (TIN) data instead of gene expression. TINs have been proposed to measure RNA integrity based on the uniformity of the read distribution across transcript length[15]. TIN-based predictions of individual PMI have similar accuracy to those based on gene expression (Supplementary Figure 36a and 36b). However, there is only moderate intersection between the two methods on the genes contributing the most to the predictions (Supplementary Figure 36c). This is consistent with our finding that the post-mortem transcriptomic changes are both the result of RNA degradation and of regulated gene expression.

Discussion

Here we report on the largest systematic study of the impact of death and post-mortem cold ischemia on gene expression across multiple human tissues. Samples obtained post-mortem are a valuable source of material for studies requiring organs and tissues difficult to obtain, or those where it is impossible to study and manipulate them in living organisms. Hence, understanding the impact of death in tissues is essential for the proper interpretation of post-mortem gene expression levels as a proxy for in vivo, living physiological levels.

The death of an organism clearly has an immediate impact on tissue transcriptomes, as illustrated by our analysis of ante- and post-mortem samples. Changes in gene expression as a response to death, and during subsequent post-mortem ischemia, might be expected to reflect stochastic variation resulting from the

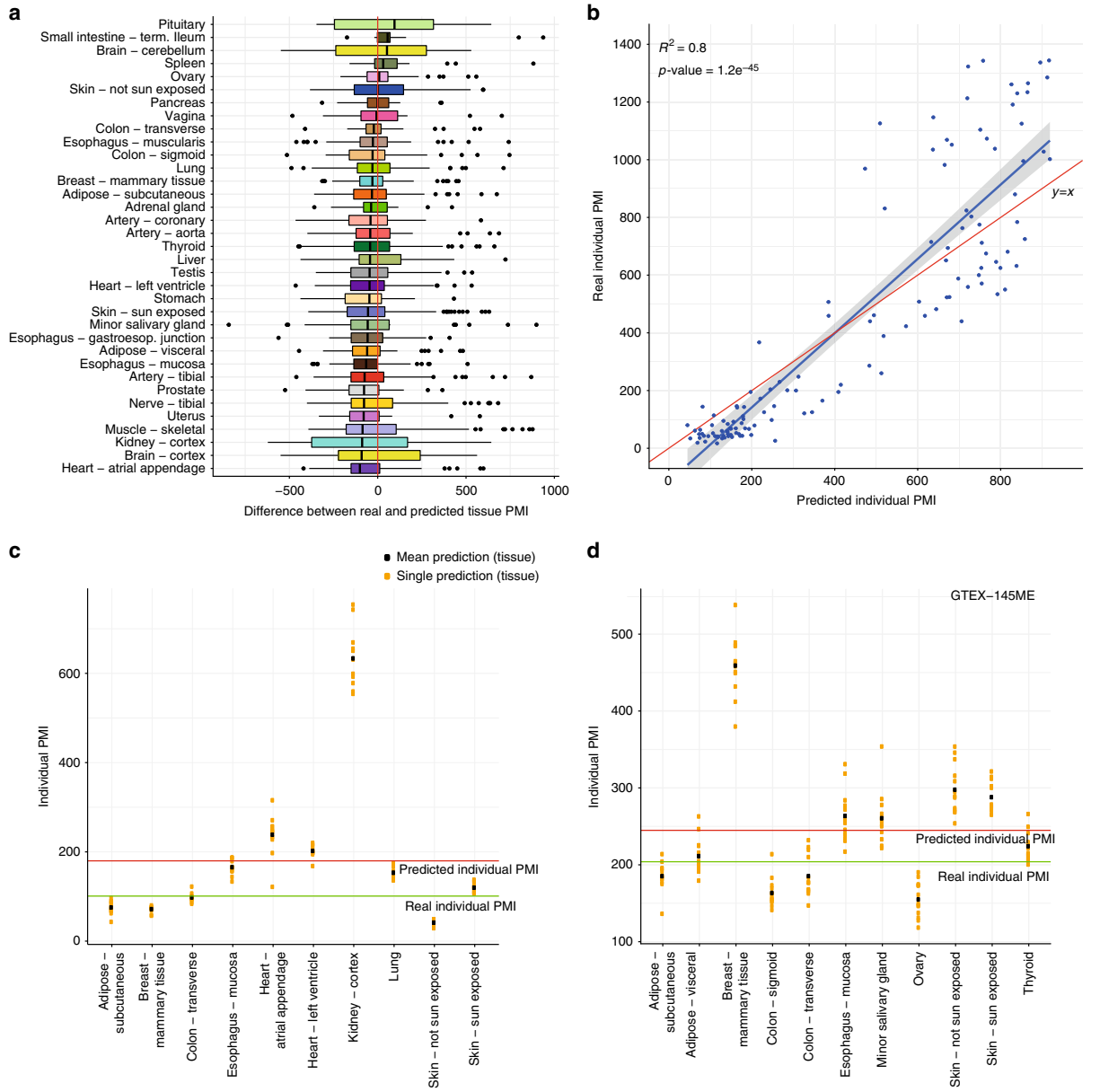


Fig. 8 Prediction of the PMI from gene expression in post-mortem samples. **a** Distribution of the PMI prediction error per tissue. **b** Regression of the real PMI versus the predicted individual PMI on the test set of 129 individuals. Plots in panels (**c**) and (**d**) illustrate two examples (GTEX-145MN and GTEX-145ME) of the prediction of the PMI for an individual based on the prediction of PMI from each tissue from the individual. For a given tissue, each yellow dot represents a prediction from each one of 13 different models. The black dot is the mean prediction of these 13 models. The green line represents the real PMI value for the individual. The individual PMI prediction is calculated as the average of the final tissue PMI predictions, and is represented by the red line

enzymatic processes underlying mRNA degradation. However, our results suggest instead that there is ongoing regulation of transcription, at least during the hours immediately following death. We observed that in the majority of tissues there are many genes that display expression profiles that are more complex than simple monotonic changes with PMI. This is in agreement with a recent study by Pozhitkov et al.[29], in which gene expression profiles produced by cDNA microarrays were analyzed in zebrafish and mouse samples with post-mortem intervals up to 48 and 96 h. That study showed a non-monotonic increase in the abundance of certain transcripts and suggested that previously silenced genes were actively transcribed at later post-mortem time points.

Similarly, analysis of splicing changes with PMI did not show conclusive evidence of systematic splicing deregulation (as measured by the splicing entropy) across tissues. Death, in contrast, did apparently lead to some splicing deregulation in blood. In particular, we found that the usage of the major isoform (the most abundant isoform compared to the rest) was attenuated in post- compared to pre-mortem samples. The usage of a major isoform

Protocol for post-mortem interval prediction of an individual

1. Annotate the start time of the forensic procedure, START_TIME.
2. Annotate the stabilization time of the tissues, among the following ones, that are available:
   - o Skin sun exposed (lower leg) sample, SKNS_STAB.
   - o Adipose Subcutaneous sample, ADPSC_STAB.
   - o Thyroid sample, THYROID_STAB.
   - o Lung sample, LUNG_STAB.
3. Prepare sample, extract RNA and perform RNA sequencing (RNA-seq) analysis, possibly with a portable and real-time equipment of the available tissues.
4. Process the RNA-seq data to obtain gene expression quantifications for each available tissue sample.
5. Provide to the PMI prediction software the gene expression values in each available tissue together with the stabilization time of the tissue, and the time of the initiation of the forensic procedure.
6. The prediction software provides tissue and individual PMIs, as well as the coefficient of variation of the tissue PMIs as a measure of the stability of the predictions

Fig. 9 Protocol for post-mortem interval prediction. Steps to be performed to predict the PMI of an individual

across tissues and biological conditions has been reported as a general characteristic of genes[38,39], and has lead to extensive debate on the physiological relevance of regulated alternative splicing[39]. Since most data to date has been collected from post-mortem samples, the preferential usage of a single major isoform in living cells may be even more prevalent than previously reported.

While the effects of death per se on gene expression are distinct from those of increasing PMI, we found a number of genes involved in the assembly of DNA, nucleosome and chromatin that were affected by both. Expression changes of these genes suggest a possible form of gene regulation through the alteration of the chromatin structure. Pozhitkov et al.[29] described an increased expression of epigenetic regulatory genes, hypothesizing that the activation of these genes reveals the nucleosomes and allows for the later transcription of developmental genes that have no early expression. We consistently detected the upregulation of genes involved in DNA organization, but we could not detect changes in expression among genes related to development. This could be due to the lack of reference expression levels from very early post-mortem samples.

Based on the tissue specific response of the transcriptome to PMI, we built machine-learning models to predict the time of death of a recently deceased individual. We show that RNA-seq performed on a few key tissues could become a powerful tool to aid in forensic pathology. It could carry the footprint not only of the time since death, but also of the cause of death—even though we could not properly carry out these analyses because of the small sample sizes available. Interestingly, the most informative tissues to predict time of death included readily accessible ones, such as skin and subcutaneous adipose. While these results show promise, larger datasets more balanced across a wider post-mortem time interval will be required to assess the full potential of the approach.

In line with previous studies[12,13,15,17,20] our analyses show that the investigation of the impact of post-mortem ischemia in tissue transcriptomes is essential to properly interpret gene expression estimates obtained from post-mortem tissue samples. Furthermore, understanding the transcriptional changes occurring with time after death could have multiple applications. Here, we illustrated an application specific to forensic pathology, but other applications could include improving biospecimen procurement and organ preservation protocols. These could, in turn, have an impact on the procedures employed for organ transplantation.

Post-mortem tissues are irreplaceable sources for researching and understanding human biology, but as our study demonstrates, death does introduce a bias in the cellular transcriptomes, even over relatively short timeframes. Ideally cellular transcript levels should be measured "in vivo" in unperturbed cells to provide an unbiased characterization of true physiological cellular transcriptomes, and this should in turn be performed individually in each of the millions of cells that constitute a living tissue. However, current technologies for the genome-wide characterization of the transcriptome still require the dissociation and destruction of cells, even when obtained from living donors, and the impact of this cellular destruction on the transcriptome is also largely unknown. Despite this, our results indicate that overall, relatively few genes show significant changes over the post mortem intervals studied, and the genes that do change do not change systematically, but vary by cell and tissue type. To minimize and limit the impact of these changes, adhering to strict protocols and standards for the collection of high quality tissue and RNA, combined with careful documentation of all key sample procurement covariates (as was done for GTEx[21]), can allow the effects of post-mortem ischemia to be largely identified and corrected for in analyses. Post-mortem samples can therefore be of tremendous value for studies of both normal and disease biology.

Methods

Data and filters. We used mRNA sequencing data (Illumina paired-end, 76 bp) from the GTEx project Analysis Freeze V6 release (phs000424.v6.p1). RNA-seq libraries are non-strand specific, with Poly-A selection and generated with Illumina TruSeq protocol. Further details on sample collection, processing and quality control of the RNA-seq samples from version V6 can be found in the supplementary material of[24] and in[21,23,40]. As described in Carithers et al.[21] the GTEx project made an effort to collect tissues within 8 h of PMI and RIN values ≥6. All samples under analysis in this study were collected and preserved with the PAXgene Tissue preservation system developed by Qiagen[21].

From the 55 available tissues in the V6, we started by selecting those with at least 20 samples. Brain samples are preserved either with PAXgene Preserved or Fresh Frozen methods. The latter does not have ischemic time available. We have included 161 samples from Cerebellum and 147 from Cortex preserved with PAXgene method and with ischemic time available. We further removed cell lines from the set of tissues. The final dataset comprises 36 tissues, with 36–1049 samples with a mean of 253 samples, see Supplementary Fig. 1.

Gene expression. RNA-seq reads were aligned to the human genome (hg19/GRCh37) using TopHat[41] (v1.4) and Gencode annotation v19[26] was used for gene quantification. We considered genes with at least 5 reads mapping in exons and from all biotypes in the annotation. The raw read counts were used for differential expression analysis and the RPKM[42] values, which were log2 transformed with an added pseudo-count used in the remaining analysis. For the regression analyses, the matrix of expression values was obtained for the samples of each tissue and then normalized with the normalize.quantiles function from the preprocessCore library[43].

In order to investigate the global patterns of gene expression we have considered the gene expression values of all the tissues and all the genes from the annotation[26]. We have then performed multi-dimensional scaling (MDS) using the isoMDS function from the package MASS in R. We defined the distance for two samples A and B as:

dist(A, B) = 1−PearsonCorrel(A, B). As shown in Supplementary Fig. 3 there is a clear transcriptional signature characterizing each tissue. Further discussion on detectable and tissue specificity expression and gene expression patterns across tissues can be found here[25].

Post-mortem interval (PMI) information. GTEx annotation provides information on three types of ischemic time: Total Ischemic time for a sample, Total Ischemic time for a donor and Ischemic Time (time for the start of the GTEx procedure). All these variables are quantified in minutes. Throughout the text we have used the term Post-Mortem Interval (PMI) to refer to ischemic time and except if explicitly stated it refers to the sample ischemic time. Negative PMI values (observed in blood samples) correspond to samples extracted pre-mortem. GTEx annotation contains a total of 249 sample and subject variables, which are divided in six groups: Sample attributes (prefix SM), Death circumstance (DTH), Demographic (no specific prefix), Medical History (MH), Tissue Recovery (TR), Serology results (LB). We selected those variables with > = 2 and < = 15 values, removing cases of unknown values. We then computed a linear regression with PMI, obtaining the adjusted R^2 and the Pearson correlation with PMI (cor.test in R with use = "na.or.complete"). Categorical variables were converted to numeric. Supplementary Fig. 4 shows the respective correlation values for all the variables, where we observe that TR, LB and DTH variables are highly correlated with ischemic time. This basically reflects different aspects of the tissue collection procedure that are highly associated with PMI.

Covariate selection. In order to assess the impact of PMI on gene expression we need to account for the possible effect of other pre-mortem variables on the variation of gene expression. For the selection of the variables of interest we excluded TR, LB and DTH variables due to their strong association with PMI, which may result from the fact that all these variables capture the underlying characteristics of death circumstances and tissue extirpation procedure. We then focused on sample, demographic and medical history variables (correlation values with ischemic time ignoring missing values), summing 166 variables. We further filter for those covariates that are qualitative and describe a phenotype such as age or gender and that exclude those that are simply metrics from the sequencing (SM). Finally, we kept those variables with $|r| > 0.1$ with PMI. Non-numerical variables are converted to numeric format. The final set of covariates used for regression analysis of PMI and gene expression is presented in Supplementary Table 5.

Gene expression and PMI regression model. In order to assess the impact of PMI on gene expression we took into account a set of fourteen covariates (Supplementary Table 5). Then, for each gene (with average expression across the tissue samples greater than 0.5 RPKM) we implemented a linear regression model where the gene expression profile is modeled with relation to the covariates: reg = lm (gene_expression ~ matrix.selectedCovariates). The residuals of the model are then used as the expression phenotype: gene_expression.resid = residuals(reg). Finally, the correlation between the residuals and the PMI is calculated, r = correlation (gene_expression.resid, pmi.vals). The corresponding correlation and p-values (adjusted with BH method[44]) are then stored for all genes. This procedure is repeated for all tissues (Supplementary Fig. 37). To compute the correlation values of gene expression and PMI without the covariates, a procedure similar to the above was used where Pearson correlation is obtained between gene expression and PMI values (Supplementary Fig. 38). Supplementary Note 2 provides the algorithmic details of the methodology.

Non-linear temporal differential expression. In order to find non-linear differential expression we developed a method to identify significant changes between different post-mortem intervals. For each tissue we grouped the samples as in[21] and in five different PMI intervals I1: < 1 h, I2: ≥ 1 h and < 4 h, I3: ≥ 4 h and < 6 h, I4: ≥ 6 h and < 15 h and I5: ≥ 15 h. We then normalized the gene expression of each gene computing a Z-score = ((X-mean)/stdev) and calculated the median expression in each of these intervals. Every two consecutive intervals, with a minimum number of five samples are then compared. We consider an event of temporal differential expression between T_i and T_{i+1}, where T_i and T_{i+1} correspond to the expression values of the gene in the interval i and $i + 1$ if we meet the two following conditions:

$$\mathrm{pval}(i,\ i+1) = \mathrm{wilcox.test}(T_i,\ T_{i+1}), \text{with pval}<0.05 \quad (1)$$

$$\mathrm{fold\_change}(i,\ i+1) = log_2(\mathrm{median}(T_i)/\mathrm{median}(T_{i+1})), |\mathrm{fold\_change}(i,\ i+1)|>2 \quad (2)$$

Supplementary Fig. 39 provides the algorithmic details of the methodology.

Tissue similarity for PMI correlated expression. In order to build a tissue similarity matrix of the correlation profiles (gene expression and PMI) we performed a pairwise comparison of all tissues. For every pair of tissues we obtain the common genes by intersecting genes that in both tissues have correlation value of gene expression with PMI. For every pair of tissues we then obtain a Spearman ranking correlation based on the correlation values of the common genes. We then used the heatmap.2 function from gplots to calculate the heatmap with dendrogram in Fig. 2f.

Functional enrichment analysis. For functional enrichment analysis we used the R libraries: DOSE[45], ClusterProfiler[46], Kegg.db[47] following the tutorial of ClusterProfiler[46].

Differential expression in blood samples. For differential expression analysis we used the statistical methods implemented in the edgeR package[48]. We started by building a matrix with gene read counts in premortem (n = 169) and postmortem (n = 223) Blood samples. Genes were filtered to have at least 5 reads per million mapped reads in at least 10% of the samples on one of the tested groups (cpm function). We created a design matrix taking into account 2 groups (pre- and postmortem samples) and several covariates:

$$\mathrm{design} <- \mathrm{model.matrix}(\sim \mathrm{SMRIN} + \mathrm{AGE} + \mathrm{ETHNCTY} + \mathrm{MHCANCERNM} + \mathrm{SMCENTER} + \mathrm{SMTSTPTREF} + \mathrm{SMNABTCHT} + \mathrm{GENDER} + \mathrm{group}, \mathrm{covars.matrix})$$

Covariates SMRIN and AGE were discretized according to the following intervals: SMRIN = < 7*/7–8/8–9/9–10 and AGE = 20–30/30–40/40–50/50–60/60–70. Covariates were converted as factors. See Supplementary Table 5 for the description of the covariates, where group variable corresponds to the pre and postmortem samples. We then followed the protocol at[48] performing the normalization with the TMM method[49] Generalized Linear Model (GLM) based functions to estimate common dispersion and differential tests. For the differential expression analysis across different post-mortem blood intervals, we first divided the postmortem samples in four groups, which provided an equal number of samples in each group: G_1 (n = 56): 0 < pmi < = 406 min; G_2 (n = 56): pmi > 406 and pmi < = 635; G_3 (n = 56): pmi > 635 and pmi < = 867; G_4 (n = 55): pmi > 867 and pmi < = 1401. A similar approach as the one described above was then applied to compare all the pre-mortem samples with each of the G_1, G_2, G_3, and G_4 groups. Figure 4b and Supplementary Table 8 shows the number of differentially expressed genes for the different intervals.

Transcriptional patterns of pre and post-mortem blood. In order to explore the transcriptional differences in pre and post mortem blood samples we have built the respective expression matrix based on RPKM values that were then log2 converted and normalized with normalize.quantiles function as previously described. We then performed hierarchical clustering (HC) and multidimensional scaling (MDS). We defined the distance between samples a and b as, dist(a,b) = 1 − cor(a,b), where cor is the Pearson correlation of a and b vector. Hierarchical clustering solution was then computed with hclust function using the average method. Visualization was performed using the heatmap.2 function with the input of the distance matrix and the previously calculated HC solution as the dendrogram parameter. Postmortem samples (n = 20) with a PMI smaller than the respective individual PMI were excluded. Heatmap with PMI interval colors is shown in Supplementary Fig. 20, and samples in MDS plot were colored according to Hardy Scale (Fig. 4a).

Signaling pathway models. The hiPathia[34] tool was used for the interpretation of the consequences of the combined changes of gene expression levels and/or genomic mutations in the context of signaling pathways (see Supplementary Fig. 40 and 41). Significant circuits associated to PMI were obtained by fitting a linear model and were summarized by the median value across samples per circuit and time points. Supplementary Note 3 and Hidalgo et al.[34] provide the algorithmic details of the methodology.

Gene structural features. Features were derived from the Gencode annotation v19[26], including the number of projected (non-redundant exonic regions) exons, length of the coding regions, overall length of the gene, biotype. We obtained projected exons first by sorting by genomic coordinates and then by merging exons. We used bedtools[50] for this step. GC content was obtained from the Ensembl Biomart (www.ensembl.org/biomart). For each tissue we have calculated the Pearson correlation between the vector of gene features and the respective correlation value between gene expression and PMI. Supplementary Table 7 contains the correlation values per tissue for each feature.

Mitochondrial transcription. For estimating the mitochondrial RNA concentrations (MT%), we divided all reads in annotated mitochondrial (mt) genes by the total number of reads in annotated (nuclear and mitochondrial) genes. To account for the substantial different mitochondrial activity across tissues, we divided each sample by the median MT% found in the corresponding tissue (nMT%). We then regressed a linear model nMT% ~ PMI and compared the slopes obtained at each time point between the different tissues (Supplementary Fig. 17). Correcting for the influence of age in the linear model changed the distribution of relative MT% only marginally (shifted values < = 0.05) (Supplementary Fig. 16).

RNA-seq metrics across tissues. We have explored if the different tissues show differences in RNA-seq quality control metrics obtained with the RNA-SeQC[40]

pipeline. Supplementary Table 11 lists the variables used for this analysis. The mapping proportions along the different gene features are shown in Supplementary Fig. 11 and 12. Degradation of the RNA may result in different mapping bias effects, in particular in a higher read coverage at the 3′ end of the genes. We have calculated for each sample a read coverage ratio between the 5′ and the 3′ 50bp-based normalization. Distribution of these values for the different tissues is shown in Supplementary Fig. 13 and the relation with RIN and PMI are shown in Supplementary Fig. 14 and 15.

Clustering modularity. To assess if gene expression signatures of tissues are preserved across the PMI bins, as defined above (section "Non-linear Temporal Differential Expression"), we selected only the tissues that had at least 10 samples within each bin of PMI. Because differences in the number of samples per tissues can introduce variation in the network structure, we randomly selected the same number of samples per tissue in each PMI bin, corresponding to the minimum number of samples per tissue across all the PMI bins. Thus, we have 4 combinations of 422 samples, one for each PMI bin, with the same tissues, and the same number of samples per tissue. For each combination of samples, we compute pairwise Pearson's correlation coefficient on the log2-transformed RPKM expression values after adding a pseudo-count of 1. From each matrix of correlation coefficients we built 7 networks, where nodes are the samples and edges are connections between samples that are correlated with a coefficient higher than a given threshold (out of 7 thresholds, from 0.86 to 0.92). These thresholds gave comparable network densities, defined as the proportion of connected nodes over the total number of possible edges, across all the networks. We used the modularity formula (R package igraph[51], modularity function) to measure how well the samples in each network are aggregated by tissue type. Supplementary Fig. 9 shows the distribution of modularity with relation to network density.

Exon inclusion analysis. GTEx samples were processed through the Integrative Pipeline for Splicing Analyses (IPSA) pipeline with default settings[52]. Namely, short reads were mapped to human genome (hg19/GRCh37) using TopHat[41] (v1.4). The alignments were filtered to have an overhang of at least 8 nt and entropy of the offset distribution of at least 1.5 bits. Novel short exons (shorter than the read length) were predicted using reads with more than one split with canonical GT/AG splicing nucleotides and minimum entropy of at least 1.5 bits for each splice junction. The percent-spliced-in (PSI) metric was computed as in Wang et al.[35] by using inclusion and exclusion reads with the minimum total count of 5 reads; that is, exons for which the combined number of inclusion and exclusion reads was less than 5 were excluded.

From the PSI values calculated by the IPSA pipeline[52] we have performed further filtering on a tissue basis based on the three following criteria: Select exons with NAs in less than 10% of the cases; Select exons they have a standard deviation greater than 0; Select exons if the difference between the max and min PSI values is larger than 0.1;

From this subset of selected exons we have then performed correlation analysis of the PSI value with the PMI value for each tissue. Supplementary Fig. 18a shows the number of tested exons according to the above filtering and Fig. 3d the number of significant exons at 1% FDR and $|r| > 0.5$.

Differential exon inclusion in blood. Exons with PSI values following the three criteria defined in the previous section in Blood samples were selected for differential expression analysis. This yielded a set of 3441 exons. Next, differential exon inclusion was tested with Wilcoxon Sum Rank test, with multiple testing adjustments by Benjamini-Hochberg method[44], and the median of the PSI values in each group calculated. Exons were deemed significant included if they pass the following criteria: FDR < 1%; and $|\Delta PSI| > 0.1$;

Cellular composition. In order to perform gene expression signal deconvolution we applied CIBERSORT[35] v1.04 and the LM22 gene signature to all blood samples, using gene RPKMs, with default parameters, deconvoluting the signal into 22 different cell types. We discarded four cell types with average fraction below 0.01 in both conditions, keeping 18 cell types. We compared the cell-fractions of pre- and post-mortem samples globally using the Anderson–Darling test[53] and by cell-type obtaining a p-value using the two-sided Wilcoxon Rank-Sum test[54], adjusted to multiple-testing by Benjamini–Hochberg method[44].

Splicing entropy analysis and PMI association. To investigate changes in patterns of isoform usage and how these correlate with the PMI we have calculated the splicing entropy based on the relative abundance of an isoform/transcript within a gene. The following selection criteria and calculation was performed on a tissue-by-tissue basis: start by selecting genes with two or more isoforms; next, select genes with a non-zero expression in 90% of the samples of the tissue. Calculate isoform ratios for each gene: For a gene G, with k isoforms I, the splicing ratio is defined as: $P(I_i) = \frac{I_i}{\sum_{i=1}^{k} I_i}$, where I_i corresponds to the RPKM value for the isoform i of G.

Finally, calculate the entropy of a gene based on the Shannon Entropy formula, as:

$$E(G) = -\sum_{i=1}^{k} p(I_i) \times \log p(I_i)$$

The splicing entropy of gene G is maximal if all its isoforms have the same ratio and minimal if one of the isoforms dominates all the expression of G.

Then, for each gene, we correlate the splicing entropy with the respective PMI of the sample. From this test, we obtain the r-value and p-value. Perform p-value adjustment for multiple testing by Benjamini–Hochberg method[44]. We repeat this analysis for all the selected tissues. In Supplementary Table 10 we provide the total number of genes tested per tissues and the genes with a $|r| > 0.5$ and FDR < 5%. Figures 3g, h present an example of a gene with a significant change in lung. Supplementary Fig. 19 presents the distribution of the correlation values for Splicing Entropy and PMI across the different tissues.

Machine learning models for PMI prediction. The predictive model for PMI based on gene expression was constructed with a two-step approach using an ensemble of gradient boosted trees (Supplementary Fig. 24) in order to provide a robust estimate of PMI and avoid overfitting. 528 available individuals are initially partitioned into training and testing datasets, using 75% and 25% of the data, respectively. This partition is performed in such a way that we try to keep a similar underlying distribution of the number of available tissues per individual both for the training and testing datasets.

In order to build these tissue models (with the R implementation of the xgboost package[37]), we first create a fixed split of individuals into training and test sets. For a given tissue, we perform 3-repeat-5-fold cross validation with the samples corresponding to the individuals of the training block in order to select the best model, and we generate the predictions over the unseen test set using this model. This process is repeated 13 times using different seeds to take into account the variation in the hyperparameter optimization process. The output is a matrix of n samples×13 columns, where each column represents the tissue PMI prediction of all test samples for each iteration. The final tissue PMI predictions will be taken as the row average of this matrix. Only protein coding genes with a correlation >= 0.4 with the tissue ischemic time were used in order to reduce the computational burden of model fitting. No other covariate was considered. Hyperparameter search was performed during this cross-validation loop using standard grid search for tree depth ranging from 4 to 6, η ranging from 0.001 to 0.1, γ ranging from 0 to 0.15 and up to 1000 rounds, using RMSE as optimization criteria. For each tissue we repeat the previous process 13 times using different seeds to determine the training set fold partitions in order to have a measure of the variability of the final prediction while applying the models on the test set (Supplementary Fig. 24b). On Supplementary Fig. 25a, we show the variability of the number of genes selected by the each 13 models, per tissue.

Once we have obtained the 13 models per tissue, we use each one of them to generate PMI predictions for each tissue sample in the test set. Therefore, 13 predictions are generated per tissue for a given test individual. We take the average of these predictions as our final PMI prediction for that particular tissue. On Supplementary Fig. 27 we can see examples of the tissue performance on the test set.

In the second step of our procedure, for each individual we will correct the final tissue PMI predictions by subtracting the elapsed time of the GTEx procedure. Since we know how much time has passed since the beginning of the GTEx procedure until a specific sample has been processed, this time difference has to be subtracted from the tissue PMI predictions in order to normalize them to a reference level, which is the start of the GTEx procedure. Finally, to predict the individual PMI (which is considered to be the time from death until the beginning of the procedure), we compute the average of these corrected tissue PMI predictions.

One important remark is that the final quality of the individual PMI prediction for the test individuals will highly depend on how accurate the individual tissue models are. For this reason, while performing the second step of the prediction procedure, we decided to use only the top 20 tissues with the best R^2 performance in the training data (Supplementary Fig. 25b). The R^2 for each of these tissues while applying the models on the test set is shown on Supplementary Fig. 26. The density of the individual PMI prediction error, which is defined as the signed difference of the real and predicted individual PMI is shown on Supplementary Fig. 30.

To investigate whether we are recovering a real predictive signal from gene expression instead of just an artifact, we used whole Blood samples, where there is information available for pre-mortem individuals as well. We reasoned if we were able to predict accurately the time to death for pre-mortem individuals this would reflect overfitting. To this end, for each cohort (pre-mortem and postmortem), we partitioned the data into training and testing datasets, fitted the model on the training data with 3-repeat-5-fold cross validation, performed the predictions on the test set and then obtained the regression statistics of real vs. predicted Blood PMI. We repeated this process a hundred times with different seeds, generating a different training/testing partition each time, in order to study the variability of the regression statistics. There is a significant difference of the regression statistics

between pre-mortem and postmortem samples, with a very poor fit for pre-mortem samples (Supplementary Fig. 29).

To study the stability of the tissue PMI predictive models we decided to evaluate the tissue performance irrespective of the training/testing partition used so far. For each tissue we partition all the available samples into 75% for training and 25% for testing. Once each tissue model is fitted on the training set with 3-repeat-5-fold cross validation, we calculate the *p*-value of the *F*-test (Supplementary Fig. 28), R^2 and slope (not shown) for the regression of the predicted tissue PMI in the test set versus the real tissue PMI. This process is then repeated 50 times by varying the training/testing partition in order to measure the variability of the regression statistics.

In order to find the optimal subsets of tissues for predicting an individual's PMI, we computed the individual PMI with all the possible combinations of sizes 2–6 of the available tissues for each individual in the test set. Then for each set size we keep the tissue combination that performed the best for each individual, in terms of individual PMI prediction error. With this, we can calculate the proportion of times a given tissue appeared in the optimal subset of a fixed size (Supplementary Fig. 32). We see that Adipose—Subcutaneous, Lung, Thyroid and Skin—Sun Exposed (Lower leg) are the tissues that consistently appeared on more optimal subsets of all sizes; and these tissues are also the ones that appear among the top most stable ones on the previously mentioned stability analysis. We then compute the individual PMI using all the possible combinations from size 2 to 4 using these four tissues (Supplementary Fig. 33 and 34).

As an additional description of stability of the individual PMI predictions, we have computed the standard deviation (SD) and coefficient of variation (CV) of the corrected tissue PMI predictions for each individual in our original test set, and generated the density curves of these statistics, shown on Supplementary Fig. 31, both for the top 20 tissues and the top 4 tissues of the best subset analysis.

Supplementary Fig. 35 shows the distribution of the death classes in the test set individuals of our PMI prediction methodology. In order to inspect if there is any immediate effect of the cause of death on the individual predictions, we have grouped the death classes in three larger categories: cerebrovascular disease, heart disease (which groups "Ischemic heart disease" and "Other forms of heart disease") and others (which groups the remaining classes). We observe that the performance of the predictions in terms of R^2 is very similar among the death classes.

To perform the prediction using TIN[15] data, we have employed the same methodology as with gene expression data, using the same initial partition of training and testing individuals. Since it is a proof of concept, we have only performed 3 repetitions of the process instead of 13 like in the gene expression method in order to reduce the computational burden. Supplementary Fig. 36a shows the performance of the model based on the TIN measure at the tissue level, while Supplementary Fig. 36b compares the predictions based on TIN with the predictions based on gene expression at the individual level. On Supplementary Fig. 36c, we show the number of most informative genes (defined as the union of genes with importance >= 0.1 across all the 13 models in the case of the gene expression model, and the union of the genes with importance >= 0.1 across the 3 models in the case of TIN, with "importance" being a measure computed by xgboost) with respect to each tissue and data type.

Data Availability. All data are available from dbGaP (accession phs000424.v6.p1) with multiple publicly available data views available from the GTEx Portal (www.gtexportal.org). The code can be obtained at https://public_docs.crg.es/rguigo/Papers/human_PMI_transcriptome/.

Received: 9 July 2017 Accepted: 22 December 2017
Published online: 13 February 2018

References

1. Bauer, M. RNA in forensic science. *Forensic Sci. Int. Genet.* **1**, 69–74 (2007).
2. Fitzpatrick, R. et al. Postmortem stability of RNA isolated from bovine reproductive tissues. *Biochim. Biophys. Acta* **1574**, 10–14 (2002).
3. Preece, P. et al. An optimistic view for quantifying mRNA in post-mortem human brain. *Brain Res. Mol. Brain. Res.* **116**, 7–16 (2003).
4. Catts, V. S. et al. A microarray study of post-mortem mRNA degradation in mouse brain tissue. *Brain. Res. Mol. Brain. Res.* **138**, 164–177 (2005).
5. Lee, J., Hever, A., Willhite, D., Zlotnik, A. & Hevezi, P. Effects of RNA degradation on gene expression analysis of human postmortem tissues. *FASEB J.: Off. Publ. Fed. Am. Soc. Exp. Biol.* **19**, 1356–1358 (2005).
6. Heinrich, M., Matt, K., Lutz-Bonengel, S. & Schmidt, U. Successful RNA extraction from various human postmortem tissues. *Int. J. Leg. Med.* **121**, 136–142 (2007).
7. Partemi, S. et al. Analysis of mRNA from human heart tissue and putative applications in forensic molecular pathology. *Forensic Sci. Int.* **203**, 99–105 (2010).
8. Vennemann, M. & Koppelkamm, A. mRNA profiling in forensic genetics I: Possibilities and limitations. *Forensic Sci. Int.* **203**, 71–75 (2010).
9. Fordyce, S. L., Kampmann, M. L., van Doorn, N. L. & Gilbert, M. T. Long-term RNA persistence in postmortem contexts. *Investig. Genet* **4**, 7 (2013).
10. Gonzalez-Herrera, L., Valenzuela, A., Marchal, J. A., Lorente, J. A. & Villanueva, E. Studies on RNA integrity and gene expression in human myocardial tissue, pericardial fluid and blood, and its postmortem stability. *Forensic Sci. Int.* **232**, 218–228 (2013).
11. Preece, P. & Cairns, N. J. Quantifying mRNA in postmortem human brain: influence of gender, age at death, postmortem interval, brain pH, agonal state and inter-lobe mRNA variance. *Brain. Res. Mol. Brain. Res.* **118**, 60–71 (2003).
12. Feng, H., Zhang, X. & Zhang, C. mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA-sequencing data. *Nat. Commun.* **6**, 7816 (2015).
13. Gallego Romero, I., Pai, A. A., Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* **12**, 42 (2014).
14. Sigurgeirsson, B., Emanuelsson, O. & Lundeberg, J. Sequencing degraded RNA addressed by 3' tag counting. *PLoS ONE* **9**, e91851 (2014).
15. Wang, L. et al. Measure transcript integrity using RNA-seq data. *BMC Bioinf.* **17**, 58 (2016).
16. Schroeder, A. et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *Bmc. Mol. Biol.* **7**, 3 (2006).
17. Jaffe, A. E. et al. qSVA framework for RNA quality correction in differential expression analysis. *Proc. Natl Acad. Sci. USA* **114**, 7130–7135 (2017).
18. Sampaio-Silva, F., Magalhaes, T., Carvalho, F., Dinis-Oliveira, R. J. & Silvestre, R. Profiling of RNA degradation for estimation of post mortem interval. *PLoS ONE* **8**, e56507 (2013).
19. Birdsill, A. C., Walker, D. G., Lue, L., Sue, L. I. & Beach, T. G. Postmortem interval effect on RNA and gene expression in human brain tissue. *Cell Tissue Bank* **12**, 311–318 (2011).
20. Searle, B. C., Gittelman, R. M., Manor, O. & Akey, J. M. Detecting Sources of Transcriptional Heterogeneity in Large-Scale RNA-Seq Data Sets. *Genetics* **204**, 1391–1396 (2016).
21. Carithers, L. J. et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv. Biobank* **13**, 311–319 (2015).
22. GTEx Consortium, T. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
23. GTEx Consortium, T. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
24. GTEx Consortium, T. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
25. Mele, M. et al. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
26. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
27. Koppelkamm, A., Vennemann, B., Lutz-Bonengel, S., Fracasso, T. & Vennemann, M. RNA integrity in post-mortem samples: influencing parameters and implications on RT-qPCR assays. *Int. J. Leg. Med.* **125**, 573–580 (2011).
28. Musella, V. et al. Effects of warm ischemic time on gene expression profiling in colorectal cancer tissues and normal mucosa. *PLoS ONE* **8**, e53406 (2013).
29. Pozhitkov, A. E. et al. Tracing the dynamics of gene transcripts after organismal death. *Open Biol* **7**, 160267 (2017).
30. Heinrich, M., Lutz-Bonengel, S., Matt, K. & Schmidt, U. Real-time PCR detection of five different "endogenous control gene" transcripts in forensic autopsy material. *Forensic Sci. Int. Genet.* **1**, 163–169 (2007).
31. Houseley, J. & Tollervey, D. The many pathways of RNA degradation. *Cell* **136**, 763–776 (2009).
32. Yang, E. et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–1872 (2003).
33. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
34. Hidalgo, M. R. et al. High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget* **8**, 5160–5178 (2017).
35. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
36. Mirakovits, J. A. Sa. K. *Forensic Science: The Basics*. (CRC Press, Boca Raton (FL), 2010).
37. Tianqi Chen, C. G. In Proceedings of 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 785–794 (2016).
38. Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* **14**, R70 (2013).
39. Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem. Sci.* **42**, 98–110 (2017).
40. DeLuca, D. S. et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
41. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

42. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
43. Bolstad, B. M. preprocessCore: a collection of pre-processing functions (2016).
44. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
45. Yu, G., Wang, L. G., Yan, G. R. & He, Q. Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**, 608–609 (2015).
46. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
47. Carlson, M. KEGG.db: A set of annotation maps for KEGG (2016).
48. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
49. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
51. Csárdi, G. igraph: Network Analysis and Visualization. v.1.1.2. Available at https://cran.r-project.org/web/packages/igraph/index.html (2017).
52. Pervouchine, D. D., Knowles, D. G. & Guigo, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29**, 273–274 (2013).
53. Stephens, M. A. Edf statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* **69**, 730–737 (1974).
54. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1**, 80–83 (1945).

Acknowledgements

We acknowledge and thank the donors and their families for their generous gifts of organ donation for transplantation and tissue donations for the GTEx research study. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). We thank Anabela Nunes for help on image edition. This work was supported by the following grants and contracts: 1) From the US NIH: Contract HHSN261200800001E (Leidos Prime contract with NCI); Contracts 10XS170 (NDRI), 10XS171 (Roswell Park Cancer Institute), 10 × 172 (Science Care Inc.), and 12ST1039 (IDOX); Contract 10ST1035 (Van Andel Institute); Contract HHSN268201000029C (Broad Institute); R01 DA006227-17 (U Miami Brain Bank) 2) Ipatimup and i3s are partially funded by the Portuguese Foundation for Science and Technology (FCT); FEDER funds through the COMPETE 2020—Operacional Programme for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT/MCTES in the framework POCI-01-0145-FEDER-007274; 3) NORTE-01-0145-FEDER-000029, supported by NORTE 2020, under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF); 4) FCT Fellowships SFRH/BPD/89764/2012 to PO, PD/BD/128007/2016 to AS; Salary support to PGF by POPH—QREN Type 4.2, European Social Fund and MCTES, program Investigador FCT, IF/01127/2014. 5) BIO2014-57291-R from the Spanish Ministry of Economy and Competitiveness and "Plataforma de Recursos Biomoleculares y Bioinformáticos" PT13/0001/0007 from the ISCIII, and EU H2020-INFRADEV-1-2015-1 ELIXIR-EXCELERATE (ref. 6,676559) Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017' and CERCA Programme/Generalitat de Catalunya, 7) Ministerio de Educación, Cultura y Deporte, under the FPU programme (Formación de Profesorado Universitario) with pre-doctoral fellowship FPU15/03635 to MMA. 8) Award Number 1R01MH101814 and 3R01MH101814-03S1 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health. European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement RNA-MAPS_294653, Ministerio de Economía, Industria y Competitividad, referencia MEIC BIO2015-70777-P y los Fondos Estructurales de la Unión Europea (FEDER) and ISCIII del Ministerio de Economía y Competitividad, referencia PT13/0001/0021 cofinaciada por el Fondo Europeo de Desarrollo Regional (FEDER). 9) MS received funding from CNPq (grant 310132/2015-0 and 427541/2016-6), and from FAPERJ (grant E_06/2015).

Author contributions

P.G.F. and R.G. conceived the study in coordination with K.A. M.M.A., F.R., and R.G. developed the PMI prediction models; C.P.S. and M.S. analyzed the mitochondrial transcription; A.B. the clustering modularity; A.A., M.R.H., J.C.C., C.C. and J.D. the signaling pathways; R.S. and D.D. the PSI calculation; J.C. and R.A. the signal deconvolution; A.S. contributed to the differential expression analysis; P.G.F. the remaining analysis. R.N., F.A., C.O., and P.O. contributed with suggestions to the analysis. P.G.F. and R.G. drafted the manuscript with contributions from M.M.A., F.R., J.D., M.S., and K. A. All authors revised the paper.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-02772-x.

Competing interests: P.G.F., C.O., and P.O. are partners of Bioinf2Bio. The remaining authors declare no competing financial interest.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Pedro G. Ferreira[1,2], Manuel Muñoz-Aguirre[3,4,5,6], Ferran Reverter[3,4,5,7], Caio P. Sá Godinho[8], Abel Sousa[1,2,9], Alicia Amadoz[10], Reza Sodaei[3,4,5], Marta R. Hidalgo[11], Dmitri Pervouchine[3,4,5,12], Jose Carbonell-Caballero[13], Ramil Nurtdinov[3,4,5], Alessandra Breschi[3,4,5], Raziel Amador[3,4,5], Patrícia Oliveira[1,2], Cankut Çubuk[11], João Curado[3,4,5], François Aguet[14], Carla Oliveira[1,2], Joaquin Dopazo[10,11,15,16], Michael Sammeth[8], Kristin G. Ardlie[14] & Roderic Guigó[3,4,5]

[1]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Rua Alfredo Allen, 208, Porto, 4200-135, Portugal. [2]Institute of Molecular Pathology and Immunology, University of Porto, Rua Dr. Roberto Frias s/n, Porto, 4200-625, Portugal. [3]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, Barcelona, E-08003 Catalonia, Spain. [4]Universitat Pompeu Fabra (UPF), Barcelona, E-08003 Catalonia, Spain. [5]Institut Hospital del Mar d'Investigacions Mediques (IMIM), Barcelona, E-08003 Catalonia, Spain. [6]Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, E-08034 Catalonia, Spain. [7]Universitat de Barcelona, Barcelona, E-08028 Catalonia, Spain. [8]Institute of Biophysics Carlos Chagas Filho (IBCCF), Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, 21941-902, Brazil. [9]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, CB10 1 SD, UK. [10]Department of Bioinformatics, Igenomix S.A, Valencia, 46980, Spain. [11]Clinical Bioinformatics Area,

Fundación Progreso y Salud (FPS), Hospital Virgen del Rocio, Sevilla, 41013, Spain. [12]Skolkovo Institute of Science and Technology, 100 Novaya Street, Skolkovo, Moscow Region 143025, Russia. [13]Chromatin and Gene expression Lab, Gene Regulation, Stem Cells and Cancer Program, Centre de Regulació Genòmica (CRG), The Barcelona Institute of Science and Technology, PRBB, Barcelona, 08003, Spain. [14]The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [15]Functional Genomics Node (INB), FPS, Hospital Virgen del Rocio, Sevilla, 41013, Spain. [16]Bioinformatics in Rare Diseases (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), FPS, Hospital Virgen del Rocio, Sevilla, 41013, Spain. Manuel Muñoz-Aguirre and Ferran Reverter contributed equally to this work.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

A limited set of transcriptional programs define major cell types

Through bulk RNA-seq gene expression, it has been previously shown that tissues and organs have characteristic transcriptional profiles [106]. These profiles, however, are the result of an heterogeneous cell type mixture, with each cell type contributing differently to the expression of each gene. In this work, we produced 53 human primary cells from 10 anatomical sites and together with GTEx data, found that cells in the human body cluster into five major cell types: epithelial, endothelial, mesenchymal, blood and neural. These have a close correspondence with the basic histological types that are used to classify tissues. We observe that these major cell types are independently replicated in mouse single-cell gene expression, and hypothesize that the underlying transcriptional programs could generalize to mammals. To bridge human bulk tissues transcriptomes with transcriptional identity at the cell type level, we derive gene signatures for these five major cell types and computationally infer their enrichments on human tissues samples. We find that tissue identity based on the abundance of the five major cell type replicates the tissue identity at the gene expression level. By integrating transcriptomics, free-text processing of pathology reports and image analysis, we validate the estimations of celullar enrichments and demonstrate how departures from normality in these enrichments correlate with histological phenotypes associated with disease.

Summary of my key contributions

- Analyzed single-cell RNA-seq data from the *Tabula Muris* project for 20 mouse organs and tissues, and found that most individual cells (and cell types) also cluster together into the five major cell types discussed in this work, suggesting the possibility of the existence of a limited number of core transcriptional programs in mammals in general and not only in humans.

- Identified histological phenotypes from nearly 8000 free-text pathology reports associated with GTEx histological images using part-of-speech tagging, fuzzy string search and parse trees. These histological phenotypes are now integrated (including all samples in the final GTEx release) in the GTEx histological image database.

- Benchmarked cellular deconvolution (constrained least squares, support vector machine-based proportion estimation, CIBERSORT, reference free deconvolution) and enrichment (xCell) methods to estimate cellular abundances from bulk RNA-seq data.

- Generated cellular enrichments in 8,527 GTEx samples for the five major cell types (endothelial, epithelial, mesenchymal, neural and blood) using cell type signatures derived from 54 primary cells from the ENCODE project and from the GTEx project. Compared these enrichments with the proportions estimated by deconvolution methods. Even if deconvolution and enrichment are based on different assumptions and modelization of the problem, the estimated cell type abundance trends were mostly replicable among methods.

- Using xCell cellular enrichments, we found that:

    1. Cellular composition constitutes a characteristic signature of tissues and reflects their histological features.

    2. Departures from normal cellular composition correlate with histological phenotypes associated with disease.

- Performed dimensionality reduction (t-SNE and UMAP) of 8,527 GTEx samples using the estimated cellular enrichments for the five major cell types, and verified that these enrichments recapitulate tissue type as strongly as the clustering based solely on transcriptional profiles.

- Trained a support vector machine classifier using histological images on stomach tissue to identify the presence of mucosa and muscularis layers, and applied it over images of colon tissue which has similar histology to that of stomach, observing that the layer predictions match the differences observed at the transcriptional level with respect to cellular composition of epithelial and mesenchymal major cell types.

- Performed histological image analysis to estimate the proportion of adipocytes with respect to the total area of the tissue in WSIs of adipose tissue, observing the proportion ($\sim 84\%$) is likely to explain the endothelial cell type enrichment observed for this tissue.

- Analyzed the pathology reports of skeletal muscle tissue to derive a numerical estimation of the proportion of fat in atrophic samples.

- Generated cellular enrichment estimations for the five major cell types for samples from 19 cancer types using gene expression from the Cancer Genome Atlas Pan-Cancer Analysis of Whole Genomes Project (PCAWG). We used these to find alterations in cellular composition in cancer when compared to normal samples.

Resource

A limited set of transcriptional programs define major cell types

Alessandra Breschi,[1,2,3,9] Manuel Muñoz-Aguirre,[1,4,9] Valentin Wucher,[1,9] Carrie A. Davis,[5] Diego Garrido-Martín,[1,2] Sarah Djebali,[1,2,6] Jesse Gillis,[3] Dmitri D. Pervouchine,[1,7] Anna Vlasova,[8] Alexander Dobin,[5] Chris Zaleski,[5] Jorg Drenkow,[5] Cassidy Danyko,[5] Alexandra Scavelli,[5] Ferran Reverter,[1,2] Michael P. Snyder,[3] Thomas R. Gingeras,[5] and Roderic Guigó[1,2]

[1]*Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, E-08003 Barcelona, Catalonia, Spain;* [2]*Universitat Pompeu Fabra (UPF), E-08003 Barcelona, Catalonia, Spain;* [3]*Department of Genetics, Stanford University, Stanford, California 94305, USA;* [4]*Universitat Politècnica de Catalunya. Departament d'Estadística i Investigació Operativa, 08034 Barcelona, Catalonia, Spain;* [5]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11742, USA;* [6]*Institut National de Recherche en Santé Digestive (IRSD), Université de Toulouse, Institut National de la Santé et de la Recherche Médicale (INSERM), Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), École Nationale Vétérinaire de Toulouse (ENVT), Université Paul Sabatier (UPS), 31024 Toulouse, France;* [7]*Skolkovo Institute for Science and Technology, Moscow, Russia 143025;* [8]*Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), 1030 Vienna, Austria*

We have produced RNA sequencing data for 53 primary cells from different locations in the human body. The clustering of these primary cells reveals that most cells in the human body share a few broad transcriptional programs, which define five major cell types: epithelial, endothelial, mesenchymal, neural, and blood cells. These act as basic components of many tissues and organs. Based on gene expression, these cell types redefine the basic histological types by which tissues have been traditionally classified. We identified genes whose expression is specific to these cell types, and from these genes, we estimated the contribution of the major cell types to the composition of human tissues. We found this cellular composition to be a characteristic signature of tissues and to reflect tissue morphological heterogeneity and histology. We identified changes in cellular composition in different tissues associated with age and sex, and found that departures from the normal cellular composition correlate with histological phenotypes associated with disease.

[Supplemental material is available for this article.]

Transcriptional profiles reflect cell type, condition, and function. In tissues and organs, they are monitored in RNA extracted from millions to billions of cells (11^6–10^9) (Haque et al. 2017), likely including multiple cell types. As a consequence, the transcriptional profiles obtained from tissue samples represent the average expression of genes across heterogeneous cellular collections, and gene expression differences measured in bulk tissue transcriptomes may thus reflect changes in cellular composition rather than changes in the expression of genes in individual cells. Single-cell RNA sequencing (scRNA-seq) has indeed revealed large cellular heterogeneity in many tissues and organs (Trapnell 2015), and the Human Cell Atlas (HCA) project (Regev et al. 2017) has been recently initiated to define all human cell types and to infer the cellular taxonomy of the human body. As a step in that direction and to bridge the transcriptomes of tissues with the transcriptomes of the constituent primary cells, and to understand how these impact tissue phenotypes, we have generated bulk expression profiles of 53 primary cell lines isolated from 10 different anatomical sites in the human body. These profiles include long- and short-strand-specific RNA-seq and RAMPAGE data (Fig. 1A; Supplemental Tables S1–S4).

[9]**These authors contributed equally to this work.**
Corresponding authors: roderic.guigo@crg.cat, gingeras@cshl.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.263186.120. Freely available online through the *Genome Research* Open Access option.

Results

Major cell types in the human body

Clustering of the primary cells based on gene expression profiles revealed a number of well-defined clusters (Fig. 1B,C; Supplemental Figs. S1, S2A,B; Supplemental Methods 1). One cluster was composed of endothelial cells; a second large cluster included a mixture of cell types: fibroblasts, stem cells, and muscle cells, among others, which we collectively termed as mesenchymal. Two smaller clusters, which clustered together, were composed of epithelial cells; finally, the melanocytes clustered separately. Almost all of the individual primary cells are assigned to the proper major cell type. The exceptions are renal mesangial cells, which have contractile properties but are classified as epithelial, and lung epithelial cells, that are classified as mesenchymal. These two cell types, however, are of embryonic origin—in contrast to the vast majority of primary cells in our study, which are adult

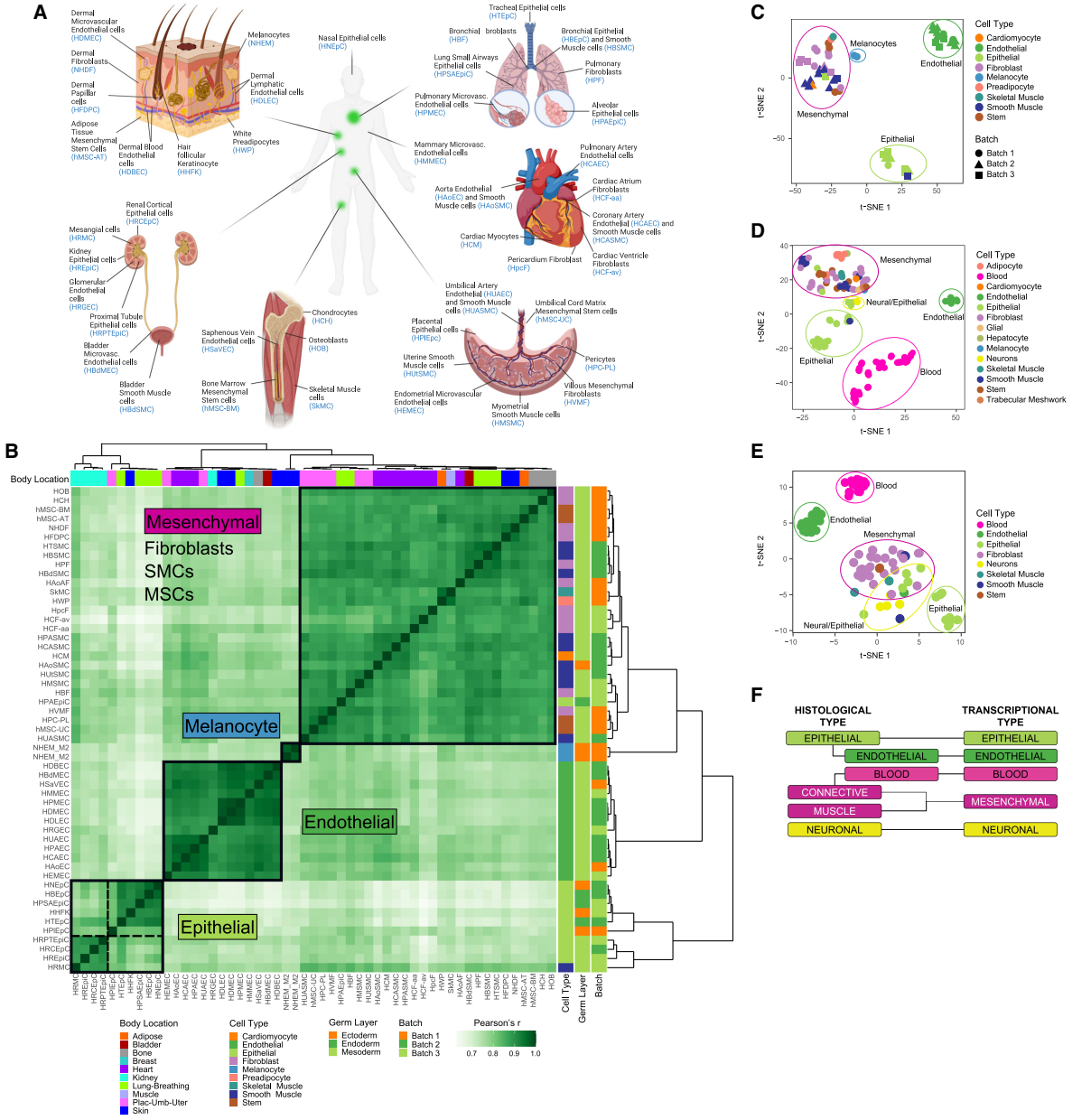


Figure 1. Basic transcriptional programs of human primary cells. (*A*) Overview of primary cells analyzed in this study and the body location they are extracted from. (*B*) Hierarchical clustering of human primary cells based on the correlation of gene expression. The clustering in four major clusters is supported by the silhouette analysis and the elbow method (Supplemental Fig. S2A,B). t-SNE of human primary cells based on gene expression measured here (*C*), on gene expression measured by CAGE by the FANTOM Consortium (*D*), and on candidate regulatory elements (cREs) by the ENCODE Encyclopedia scored DNase I hypersensitivity signal (*E*). (*F*) Correspondence between transcriptionally derived major cell types and classical histological types.

(Supplemental Table S1)—and their transcriptomes may not reflect the transcriptomes of fully differentiated cells.

The clustering of primary cells does not appear to be dominated by body location or embryological origin. Body location contributes very little to the expression profile of primary cells, explaining only ~4% of the variance in gene expression (Supplemental Fig. S2C). Variation of gene expression among organs is similar for the different clusters (Supplemental Fig. S2D). The transcriptional diversity among cells within a given organ can be as high as that across the entire human body (Supplemental Fig. S2E). A similar clustering is obtained using FANTOM CAGE-based transcriptomic data on 105 primary cells (Fig. 1D;

Supplemental Fig. S3A,B; Supplemental Table S5; Supplemental Methods 2; The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014), which reveals, in addition, two clusters corresponding to blood and neural cells, which were not represented in our set of primary cells. The analysis of a different set of primary cells from the ENCODE Encyclopedia Registry of candidate regulatory elements (cREs) (Supplemental Table S6; The ENCODE Project Consortium 2020), based on DNase I hypersensitive sites (DHSs), also recapitulates the clustering (Fig. 1E; Supplemental Fig. S3C; Supplemental Methods 3). The clustering remains in the set of 146 nonredundant primary cells that results from merging the RNA-seq, the CAGE, and the DHS data. The clustering is thus conserved despite the heterogeneity of the underlying assays and experimental protocols used to generate these different data sets (Supplemental Fig. S4). In the clustering, neural cells (mostly astrocytes from different brain regions and neurons) cluster together with a few neuroepithelial primary cells (we labeled them epithelial, but they are mostly ciliated cells from different sites in the eye). Although the neural cells profiled by CAGE seem to have a distinct transcriptional signature (Supplemental Fig. S3A), neural cells profiled by DNase-seq show a gene expression pattern similar to mesenchymal cells (Supplemental Fig. S3C). However, the neural cells profiled by DNase-seq are, in contrast to most primary cells investigated here, of embryonic origin; thus, they are unlikely to express the transcriptional program characteristic of adult neural cells. The analysis of publicly available transcriptomics data from nervous tissues, including single-cell and bulk RNA-seq, strongly support that the neural cell type is a proper major type differentiated from the other types (Supplemental Figs. S5–S7; Supplemental Methods 1).

Comparable multitissue RNA-seq data have become recently available at the single-cell level for 20 mouse organs and tissues through the *Tabula Muris* project (The Tabula Muris Consortium 2018). Principal component analysis (PCA) of the individual cells and hierarchical clustering of the primary cell types show that most individual cells, and most cell types, clustered into the aforementioned five major cell types, irrespective of the organ of origin (Supplemental Figs. S8, S9; Supplemental Methods 4). As in the case of melanocytes, we also found a few specialized cell types which do not properly belong to these types. Hepatocytes are a notable example (Supplemental Figs. S8A, S9A). Although closer to the epithelial cells than to cells of other types, they seem to have a quite specialized transcriptional program.

Altogether, these results suggest the existence of a limited number of core transcriptional programs encoded in the human genome, and likely in mammalian genomes, in general. These programs underlie the morphology and function common to a few major cellular types, which are at the root of the hierarchy of the many cell types that exist in the human body (Table 1). They all show similar transcriptional heterogeneity, with blood and epithelial within the solid tissues being the most transcriptionally diverse (Supplemental Fig. S10). These transcriptionally defined major cell types correspond broadly, but not exactly, the basic histological types in which tissues are usually classified (e.g., Eroschenko 2013; Mescher 2013; Young et al. 2013): epithelial, of which endothelial is often considered a subtype; muscular; connective, which includes blood; and neural. However, from the transcriptional standpoint, endothelial constitutes a separate type, closer, if any, to the mesenchymal than to the epithelial type. Blood is also a separate major cell type, and the connective (but not blood) and the muscular histological types cluster together into a single mesenchymal transcriptional type (Fig. 1F).

Table 1. Cell types in the human body

Cell type	Sets of cells with similar phenotype (morphology and functions). The similarity threshold induces a taxonomic hierarchy of cell types, by means of which similar cell types are recursively aggregated into higher order types.
Primary cell type	Cell types at the bottom of the taxonomic hierarchy. They denote specialized cells phenotypically identical (to some resolution); they cannot further be segregated into biologically meaningful subtypes; for example, pancreatic beta cells. In our work, we do not include cell lines here, which are primary cells that have been transformed to proliferate indefinitely.
Major cell type	Cell types at the root of the taxonomic hierarchy. They cannot be further aggregated in biologically meaningful higher order types; for example, epithelial cells.
Tissue-specific cell type	Cell type topologically restricted to a specific anatomical region (tissue, organ, body location); for example, hepatocytes.
Transcriptional program	The pattern of gene expression characteristic of a given cell type.

Within each of the major types, further hierarchical organization of cell types may exist. Although we have not profiled enough diversity of primary cells to resolve the taxonomic substructure within each major cell type, hints of this substructure can be seen in the epithelial type. Within the epithelial cluster, two well-defined subclusters can be identified (Fig. 1B–E; Supplemental Fig. S2A). One of the clusters is made mostly by renal cells, indicating that body location may play a role in subtype specialization. The epithelial cluster includes primary cells of all embryonic origins (ectoderm, endoderm, and mesoderm), suggesting that the transcriptional programs of cells may not be fully inherited through development, but partially adopted through function. The more heterogeneous composition of the epithelial type is also apparent in the mouse scRNA-seq (Supplemental Figs. S8, S9).

Our results also suggest that although many cells are likely to adhere to these basic transcriptional programs, many other primary cells are likely highly specialized and very tissue-specific. As with melanocytes and hepatocytes in our analyses, these specialized cells are likely to have their unique transcriptional program.

Cell-type-specific genes

We identified a total of 2871 genes (including 2463 protein-coding genes, 283 long noncoding RNAs, and 125 pseudogenes), the expression of which is specific to epithelial, endothelial, mesenchymal or melanocyte cell types (Fig. 2A; Supplemental Fig. S11; Supplemental Table S7). These cell-type-specific genes include nearly all genes that we identified as the major drivers of the clustering (Supplemental Fig. S12; Supplemental Methods 1). Examples of these genes include collagen (*COL1A2, COL3A1, COL6A1/A2/A3*), expressed in mesenchymal cells; epithelial transcription factors genes *OVOL1/2*; *VWF* gene encoding for the endothelial marker von Willebrand Factor; and *TYR* gene encoding for the melanocyte-specific enzyme tyrosinase (for a list of manually curated driver genes, see Supplemental Table S8). Figure 2B shows the expression pattern of *LINC01235*, an endothelial-specific long noncoding RNA (lncRNA) of unknown function. The gene is expressed in nearly all endothelial cells analyzed here, but not in cells from other types, and its expression is correlated to protein-coding genes with endothelial-related functions (Supplemental

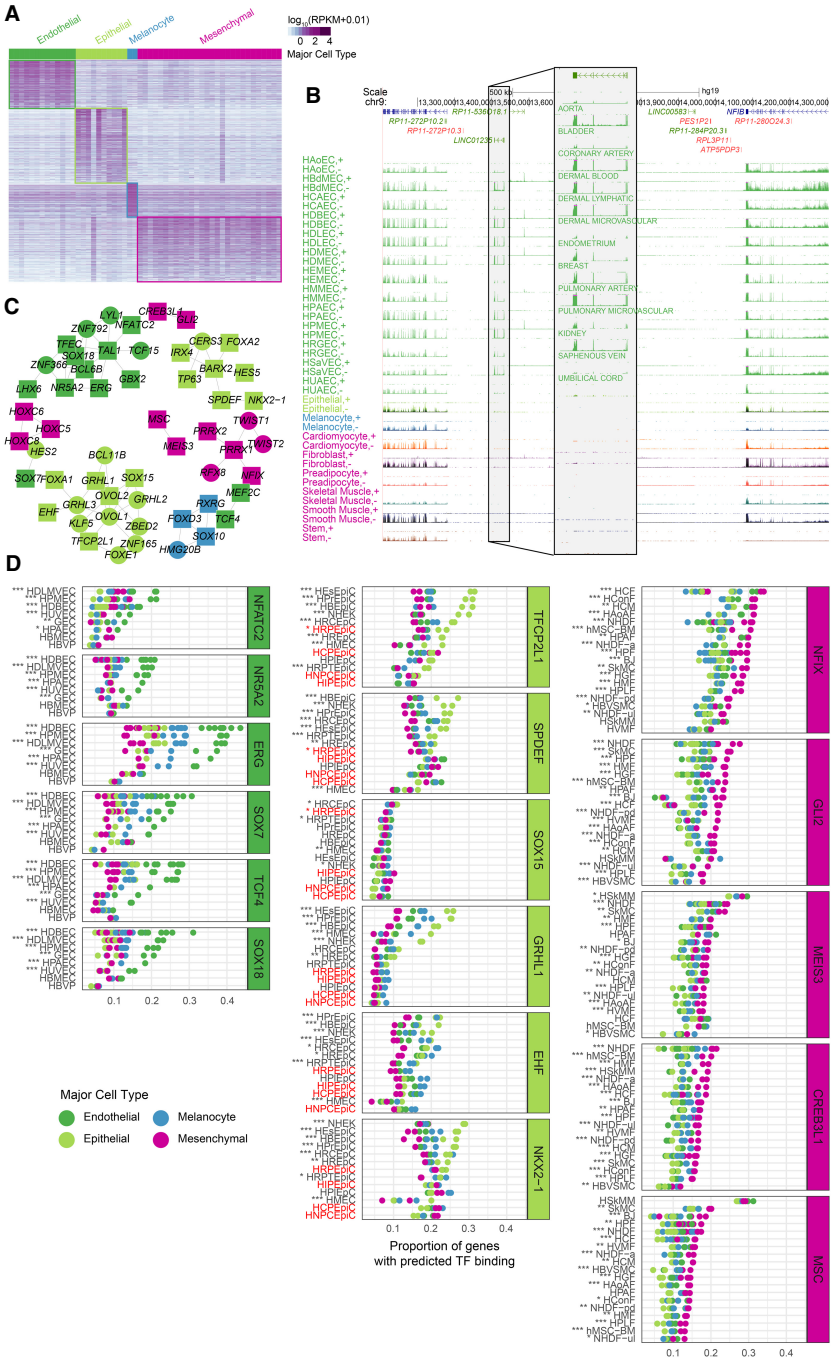


Figure 2. Cell-cluster-specific genes. (*A*) Expression of 2871 genes specific to major cell types. (*B*) Expression of the endothelial-specific lncRNA *LINC01235*. Separate strand-specific signal tracks are shown for endothelial cells, and the other tracks contain overlaid signal for each cell type. The lncRNA has highly correlated (correlation coefficient >0.9) expression with 72 protein-coding genes across our set of primary cells. Nearly all these genes are endothelial specific, and they are functionally enriched for vessel development and angiogenesis (Supplemental Fig. S13A). The gene appears to be under relatively strong regulation, because it has almost 1500 eQTLs across multiple tissues in GTEx (v7), well above the average eQTLs for lncRNAs (about 450). (*C*) Network of the most strongly coexpressed (Pearson's $r > 0.85$) cell-type-specific transcription factors (TFs). Nodes are colored according to the cell type specificity of the TF, and they shaped based on the availability of sequence motif: (square) available; (circle) not available. (*D*) Proportion of cell-type-specific genes with predicted TF binding over cell-type-specific genes that harbor a DHS around their TSS (–10 kb/+5 kb), individually for each cell-type-specific TF (with binding motif available) and cell line for which DNase-seq data was available. In general, we found that genes specific to a given type are enriched for binding motifs for TFs specific to that type. For example, the proportion of endothelial-specific genes with DHS sites that harbor motifs for the endothelial-specific TF ERG in dermal blood endothelial cells (HDBEC) is larger than the proportion of genes with DHS sites specific of other major cell types. Primary cells highlighted in red, although included within the epithelial major cell type, they have been labeled as neural/epithelial in Figure 1D, and they are therefore not proper epithelial; consistently, they do not show the enrichment in binding motifs for epithelial-specific transcription factors. Refer to Supplemental Table S6 for a complete description of the acronyms. Enrichment adjusted *P*-values: (*) <0.05; (**) <0.01; (***) <0.001.

Fig. S13A). The gene, however, is expressed in multiple tissues; therefore, it is not tissue specific.

The functions of annotated tissue-specific genes closely match the expected biology of the primary cells in each type (Supplemental Fig. S13B). Cell-type-specific genes show consistent restricted expression in the FANTOM CAGE data (Supplemental Fig. S14), and they are enriched for encyclopedia cREs (Sheffield et al. 2013) specifically in the primary cells of that type (Supplemental Fig. S15). Using ChIP-seq histone modification data obtained in a number of primary cells (Supplemental Table S9; Supplemental Methods 5; The ENCODE Project Consortium 2012), we found the promoters of genes specific to a given type to be enriched for activating chromatin marks in primary cells of that type compared with primary cells of different type (Supplemental Fig. S16A). However, overall, except for H3K4me1, we found low levels of most activating marks in the promoters of cell-type-specific genes compared with all genes, even after controlling for differences in gene expression. In contrast, the promoters of cell-type-specific genes show similar or higher levels of repressive histone modifications compared to all genes (Supplemental Fig. S16B). This is consistent with previous reports showing that genes under tighter regulation show lower levels of activating histone modifications than broadly expressed genes (e.g., Rach et al. 2011; Pervouchine et al. 2015).

Among cell-type-specific genes, we identified 167 transcription factors (TFs) from a total of 1544 TFs annotated in the human genome (Zhang et al. 2012). We focused on 56 that showed the strongest coexpression patterns (Pearson's $r \geq 0.8$) (Fig. 2C; Supplemental Fig. S17). They include previously annotated cell-type-specific transcriptional regulators, such as ERG, which has been shown to regulate endothelial cell differentiation (McLaughlin et al. 2001), and TP63, which is an established regulator of epithelial cell fate and is often altered in tumor cells (Yoh and Prywes 2015). Consistent with the hypothesis that the cell-type-specific TFs might regulate cell type specificity, we found that genes specific to a given type are enriched for binding motifs for TFs specific to that type in most cell lines (Fig. 2D). The enrichment arises specifically when the motifs occur in open chromatin domains in primary cells of that type (e.g., in epithelial primary cells, epithelial-specific genes are enriched, compared to genes specific to other types, in epithelial-specific TF motifs occurring in open chromatin domains) (Fig. 2; Supplemental Fig. S18).

We found that transcriptional regulation appears to play a major role compared to post-transcriptional (splicing) regulation,

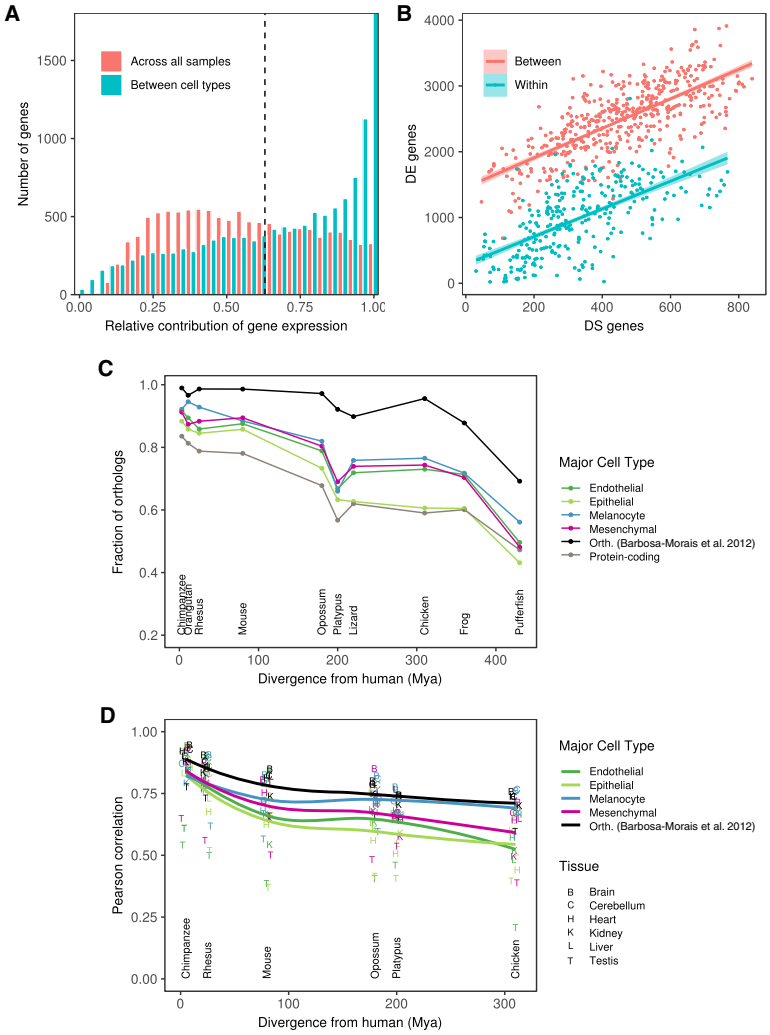


Figure 3. Transcriptional complexity of human primary cells and evolutionary conservation of cell-type-specific genes. (*A*) Distribution of the relative contribution of gene expression to the variation in isoform abundance between major cell types (blue) and between all primary cells. Large values of the contribution of gene expression indicate that changes in isoform abundance from one condition (primary cell, cell type) to another can be simply explained by changes in gene expression. Small values, in contrast, indicate that changes of isoform abundance are mostly independent of changes in gene expression and can obey changes in the relative abundance of the isoform. (*B*) Number of differentially expressed genes (DE, *y*-axis) versus the number of genes with differentially spliced exons (DS, *x*-axis), between pairs of samples of the same cell type (within, blue), or different cell types (between, red). DS genes have been obtained using IPSA (https://github.com/pervouchine/ipsa-full). See also Supplemental Figure S19. (*C*) Fraction of 1 to 1 orthologs between each species and human for major cell-type-specific genes and for protein-coding genes overall. Species are sorted by increasing evolutionary distance from human. The black line is given as a reference, and it indicates the proportion of six-way orthologs (chimpanzee, rhesus, mouse, opossum, platypus, and chicken) that are present in each species. The proportion is not 100% in these species because different versions of the GENCODE gene set reference were used. The genes in this set of six-way orthologs are used for the comparison of gene expression in Supplemental Figure S22A. See also Supplemental Figure S22C. (*D*) Pearson's *r* between gene expression in each human organ and the corresponding one in every other species. The correlation is computed across all the genes in each major cell type separately. See also Supplemental Figure S23.

both in defining the major cell types as well as the individual primary cells within the types. We estimated the fraction of the variation in isoform abundance explained by variation in gene expression (Gonzalez-Porta et al. 2012) to be on average 67% across transcriptional types and 55% across primary cells (Fig. 3A). The lower proportion of variance explained across primary cells suggests that splicing plays a comparatively more important role in defining the transcriptomes of primary cells within a given type than in setting the transcriptional programs of the major cell types. In additional support of this conclusion, we found that although the number of differentially expressed genes in pairwise comparisons of primary cells is much larger between than within cell types, the number of differentially spliced genes is similar (Fig. 3B; Supplemental Fig. S19; Supplemental Methods 6).

Although bulk gene expression is the main contributor to define cell-type specificity, other transcriptional events are also cell-type specific. First, using the RNA-seq data, we identified cell-type-specific splicing events, independent of the tissue of origin (Supplemental Fig. S20; Supplemental Table S10; Supplemental Methods 6). Second, using the RAMPAGE data, we identified cell-type-specific TSSs (Supplemental Fig. S21; Supplemental Table S11; Supplemental Methods 7).

The basic human transcriptional programs seem to have been established early in vertebrate evolution: genes orthologous of cell-type-specific genes are underrepresented compared to orthologs of all genes in invertebrate genomes (Supplemental Fig. S22A,B), but they are overrepresented in vertebrates, as early as in tetrapoda. One exception is epithelial genes, which are overrepresented only in mammals (Fig. 3C; Supplemental Fig. S22C). Within the set of orthologous genes across tetrapoda (Barbosa-Morais et al. 2012), the expression of cell-type-specific genes is less conserved than that of protein-coding genes overall, especially at larger evolutionary distances (Fig. 3D; Supplemental Figs. S22D, S23; Supplemental Methods 1). This suggests an important role in the evolution of gene expression regulation in shaping the basic transcriptional programs in the human genome. Epithelial-specific genes also show the lowest conservation of expression levels. The transcriptional program characteristic of the epithelium appears to be, therefore, the most dynamic evolutionarily—possibly reflecting a greater need for adaptation of the epithelial layer in constant interaction with the environment—and it is also consistent with the greater transcriptional heterogeneity of this major cell type.

Estimation of the cellular composition of complex organs from the expression of cell-type-specific genes

We used the patterns of expression of cell-type-specific genes to estimate the cellular composition of human tissues and organs from GTEx bulk tissue transcriptome data (version 6, 8555 samples, 31 tissues, 544 individuals) (The GTEx Consortium 2017). We used xCell (Aran et al. 2017), using the sets of genes specific to epithelial, endothelial, and mesenchymal major cell types derived from ENCODE, and specific to brain (neural) and blood derived from GTEx (Yang et al. 2018) as signatures, and computed the enrichments of these cell types in each GTEx tissue sample (Supplemental Methods 8).

The xCell enrichments (Fig. 4A; Supplemental Table S12) are largely consistent with the histology of the tissues. For example, esophagus mucosa is enriched for epithelial cells, whereas esophagus muscularis is enriched for mesenchymal cells. Skin (both exposed and unexposed) is enriched in epithelial cells and fibroblasts in mesenchymal cells. Blood and brain are only enriched in blood and neural cells, respectively. Most other tissues are not enriched in these two major cell types, with the expected exceptions of spleen enriched in blood cells and pituitary enriched in neural cells. Testis, which has widespread transcription (Soumillon et al. 2013), is also enriched in neural cells, a reflection of the similarity of the expression programs of these two organs (Guo et al. 2005). Consistent with previous observations (Frontini et al. 2012), we found enrichment of cells of endothelial type in adipose tissue. The analysis of the pathology reports of the subcutaneous adipose tissue shows that it is often contaminated with other tissues, in particular blood vessels, which would explain the enrichment in cells of the endothelial type. We have further processed and analyzed the histopathology images available from the GTEx adipose samples (Supplemental Methods 8) and estimated that, on average, ~84% of the adipose tissue corresponds to adipocytes (Supplemental Fig. S24), which would explain the endothelial enrichment. In skeletal muscle, we do not observe a particularly large enrichment in cells of the mesenchymal type, in apparent contradiction with our initial classification (Fig. 1B, F). The samples in GTEx, however, are all from differentiated skeletal muscle, whereas the ENCODE primary cells that we used to identify the mesenchymal-specific genes are undifferentiated satellite cells (SkMC) and smooth muscle cells (Supplemental Table S1). We analyzed single-cell RNA-seq data produced during skeletal myoblast differentiation (Trapnell et al. 2014) and found that differentiating skeletal muscle cells retain the mesenchymal signature through most of the differentiation pathway, acquiring only the GTEx muscle specific signature when fully differentiated (Supplemental Fig. S25A–C). Further supporting that muscle is indeed of mesenchymal type, potentially forming a well-defined subtype, gene expression profiles cluster together myoblast differentiating single cells with ENCODE mesenchymal cells, rather than with epithelial or endothelial cells, or forming a separate cluster (Supplemental Fig. S25D).

To independently assess the xCell enrichments, we analyzed the histological images of the few tissues in which samples were obtained from different subregions. These are most notable in the case of transverse colon and stomach. The GTEx stomach samples are all from the gastric body, whose walls consist of two broad layers: the mucosa, which is mostly epithelial, and the muscularis, which is smooth muscle (Fig. 4B). We processed the histological images and identified a subset of samples that presented mostly the muscularis or the mucosa layer (Supplemental Methods 8). The enrichment of epithelial cells in the samples from the muscularis layer is much lower than in the samples from the mucosa layer; conversely, the enrichment of mesenchymal cells is much higher in the muscularis than in the mucosa layer. The two sets of samples are almost perfectly separated by our cellular enrichments (Fig. 4C), explaining the bimodality in the distribution of cell type enrichments observed specifically in the stomach samples (Fig. 4A). Consistently, we found that epithelial-specific genes were exclusively expressed in the mucosa layer, and mesenchymal-specific genes were exclusively expressed in the muscularis layer (Fig. 4D). Next, we used the classification of stomach images to train an SVM model (Supplemental Fig. S26A,B) and used this model to predict the presence of the two layers in 196 transverse colon samples, with histology similar to that of stomach (Supplemental Methods 8). The SVM-predicted classification closely matches the differences observed at the transcriptional level and confirms that the

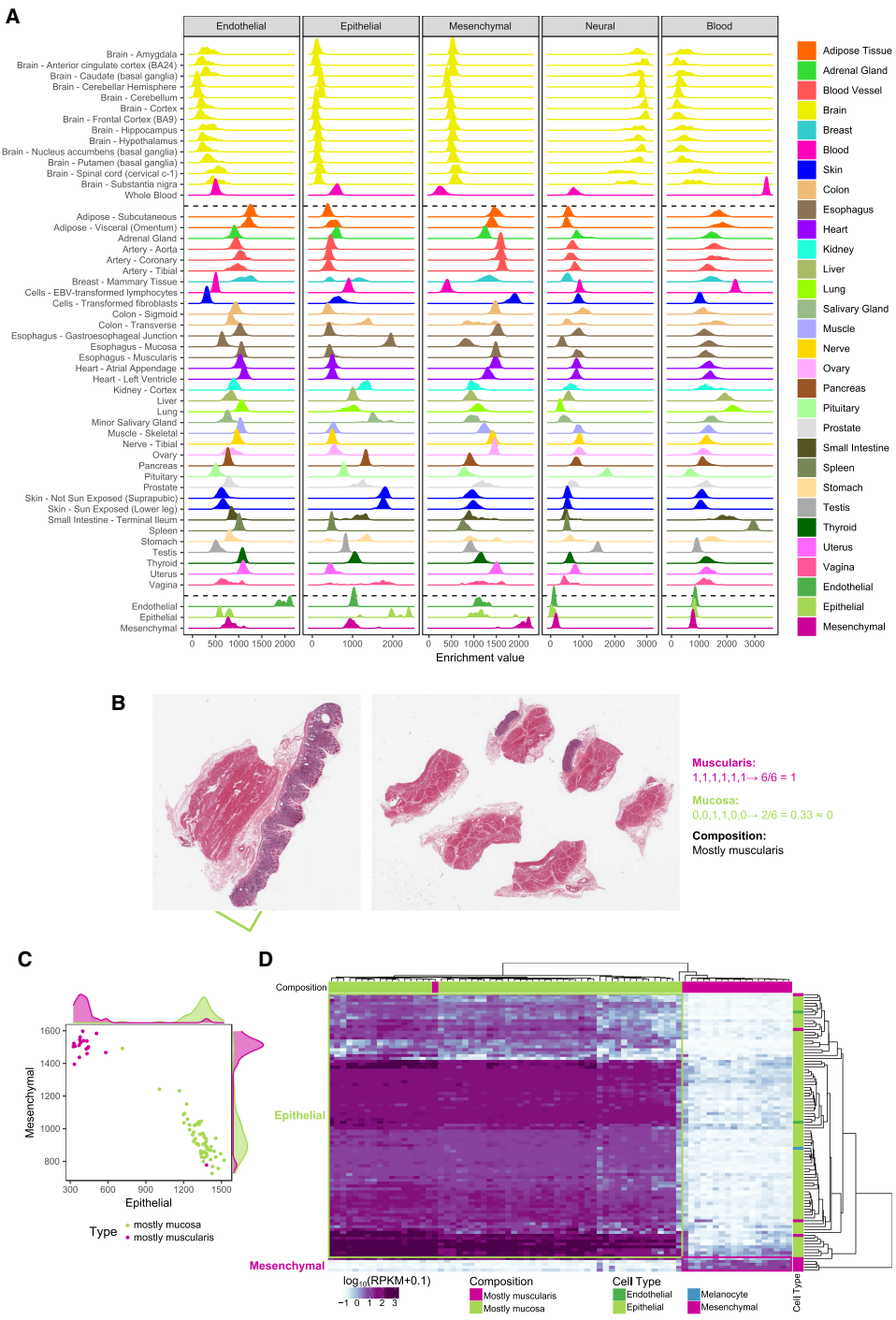


Figure 4. Expression of cell-type-cluster-specific genes in GTEx organs. (*A*) Enrichment of each major cell type in GTEx tissues, estimated from bulk tissue RNA-seq using the xCell method. As a control, we also include, at the *bottom* of the plot, the enrichments of the endothelial, epithelial, and mesenchymal primary cells monitored here (Fig. 1B). As expected, since the gene signatures have been derived from these very same cells, endothelial primary cells are heavily enriched in the endothelial type, but not in the other types, epithelial cells in the epithelial type and mesenchymal cells in the mesenchymal type. (*B*) Example of stomach histological slides, which represent the two main tissue layers and the procedure for the manual annotation of the images based on the presence of those layers. Each GTEx histological image displays up to six tissue slices. For the stomach samples, we scored each slice for the presence (1) or absence (0) of the muscularis and mucosa layers, summed up the values for each layer separately and divided by the number of slices. If the proportion of slices with mucosa layer, or muscularis layer, is more than 50% we classify the entire slide as mc1, or ms1, respectively. If the proportion is lower, we classify the slide as mc0 or ms0. A combined class, for example mc0ms1, is assigned to the slides. Thus, samples labeled mc0ms1 are mostly muscularis, and samples labeled mc1ms0 are mostly mucosa. (*C*) Enrichment of cells of epithelial and mesenchymal types in stomach samples containing mostly the mucosa (green) or mostly the muscularis (purple) layer. (*D*) Expression of the cell-type-specific genes that drive the separation of stomach samples in mostly muscularis or mostly mucosa samples. Among discriminant cell-type-specific genes, mucosa-only samples express almost exclusively epithelial-specific genes, whereas muscularis-only samples express exclusively mesenchymal-specific genes.

bimodality of cellular composition (Fig. 4A) is again related to the unbalanced presence of the two tissue layers across samples (Supplemental Fig. S26C). Considering that stomach and colon were not represented in our primary cell collection, this constitutes a strong validation of our estimates of the cellular enrichments in tissues.

Alterations of cellular composition in pathological states

We projected the solid non-neural GTEx tissue samples on a three-dimensional space according to the enrichments of epithelial, endothelial, and mesenchymal cell types in each sample (Fig. 5A; Supplemental Fig. S27). The spatial arrangement of the samples recapitulates tissue type as strongly as the clustering based on gene expression (Supplemental Fig. S28). This suggests that the basic cell type composition is a characteristic signature of tissues and that departures from this composition may reflect pathological or diseased states. To assess this hypothesis, we analyzed the histological reports associated with the GTEx images (7911 reports). We used fuzzy string search and parse trees to convert the natural language annotations produced by the pathologists to annotations in a controlled vocabulary that can be analyzed automatically (Supplemental Methods 8; Supplemental Table S13). In this way, we identified 19 histological phenotypes affecting one or more tissues for which there were at least 30 affected samples. From these, we identified six conditions with significant (FDR < 0.01) altered contributions of major cell types when comparing the composition of affected and normal tissue (Fig. 5B–E). Atherosclerosis in the tibial artery, which is more prevalent in older donors (Supplemental Fig. S29A), is associated with an increase in endothelial cells (Fig. 5B); this might be attributed to endothelial proliferation stimulated in peripheral artery occlusion (Ziegler et al. 2010). Atrophic skeletal muscle, a phenotype that is also correlated with age (Supplemental Fig. S29B), is associated with an increase in mesenchymal cells, which is consistent with the reported increase of connective tissue (Appell 1990) and intermuscular fat (Manini et al. 2007; Addison et al. 2014) in atrophy (Fig. 5C). Indeed, analysis of the pathology reports of GTEx muscle histological images reveals that the proportion of fat is almost twice as high in atrophic than in non-atrophic muscle (24% vs. 13%) (Supplemental Methods 8). Elevated enrichments of mesenchymal cells are also observed in liver congestion (Supplemental Fig. S30A), a condition that often precedes fibrosis, which is characterized by an activation of matrix-producing cells, including fibroblasts, fibrocytes, and myofibroblasts (Elpek 2014). Despite the low presence of cells of the major cell types in the testis, we found a further reduction of enrichment of endothelial cells in testis undergoing spermatogenesis (Supplemental Fig. S30B). In lung pneumonia, we also observe alteration of all cell types (Supplemental Fig. S30C). The sixth condition is gynecomastia, a pathology that is characterized by ductal epithelial hyperplasia (Cuhaci et al. 2014). We investigated differences in cellular composition between males and females and found them significant only in mammary tissue, where female breasts show much higher enrichment in epithelial cells than male breasts, possibly owing to the presence of epithelial ducts and lobules (Fig. 5D). Males diagnosed with gynecomastia show a cellular composition similar to that of females, mirroring tissue morphology.

We also observed specific age-related changes in cellular composition in lung and ovarian tissues. In lung samples we observe changes of all cell types, in particular, a significant reduction of epithelial cells in older donors (Fig. 5E), which is consistent with the impaired recellularization of lung epithelium that has been observed in decellularized lungs of aged mice (Sokocevic et al. 2013). Consistently, a similar pattern can be observed in the lungs of the individuals that died of respiratory-related causes (Supplemental Fig. S30D,E). In ovarian samples of women older than 48, a lower bound for menopause occurrence, we observe a decrease in endothelial cells (Supplemental Fig. S30F), potentially related to an age-dependent decline in ovarian follicle vascularity (Tatone et al. 2008).

Altered cellular composition is likely to be particularly relevant in cancer. Therefore, we analyzed transcriptome data from The Cancer Genome Atlas Pan-Cancer Analysis of Whole Genomes Project (PCAWG) (The Cancer Genome Atlas Research Network et al. 2013) for 19 cancers affecting tissues also profiled in the GTEx collection and estimated the cellular enrichments of the major cell types (Supplemental Fig. S31; Supplemental Methods 9). In some cases, there is also transcriptome data for normal samples from the same cancer project, which serves as a control for the highly different methodologies used in GTEx and the cancer projects. Thus, in lung cancer, there is an increase in epithelial cells (Fig. 6A,B), likely reflecting the epithelial origin of most lung cancers. In kidney primary tumors, in contrast, there is an overall increase of endothelial cells across most cancer subtypes, consistent with the increased vascularity associated with the cancer (Fig. 6C,D). The exceptions are renal papillary cell carcinomas, which instead present reduced vascularity (Aziz et al. 2013). In both cases, the cellular composition of GTEx samples and normal samples from the cancer projects are similar, supporting the robustness of our cellular characterization. Alterations in cellular composition can also reflect cancer progression. For ovary, even though we lack a comparable set of normal samples from the cancer projects, there are data on different stages of the disease, which serve as an internal control (Fig. 6E,F). Compared to GTEx normal data, there is an increase in epithelial cells in cancer, which is more evident as the severity of the cancer progresses, from primary to recurrent.

Discussion

The ultimate aim of human genetics is to understand how variations in the sequence of DNA impact organismal traits. However, the path connecting the DNA sequence of the genome to the phenotypic traits of the organism remains mostly unknown, involving a hierarchy of levels of increasing organizational complexity. This path, which unfolds during development, initiates with the transcription of DNA into RNA and its subsequent processing to functional mature RNAs. These, mostly through translation to proteins, determine cell morphology and function. Cells with similar functions aggregate to form tissues, and tissues organize into organs. Systems are made of different types of organs that work cooperatively to perform a particular function. Owing mostly to genome-wide association studies (GWASs), thousands of genetic variants have been connected to human traits and diseases. GWASs, however, uncover only statistical association. Fully understanding the causes and the mechanisms through which DNA variation impacts organismal phenotypes requires understanding how this variation impacts through each of the intermediate levels of organizational complexity. The advent of high throughput technologies to monitor transcription—microarrays first, then RNA sequencing—made possible the identification of genetic variants affecting gene expression. However, how DNA variants and the resulting molecular phenotypes propagate through

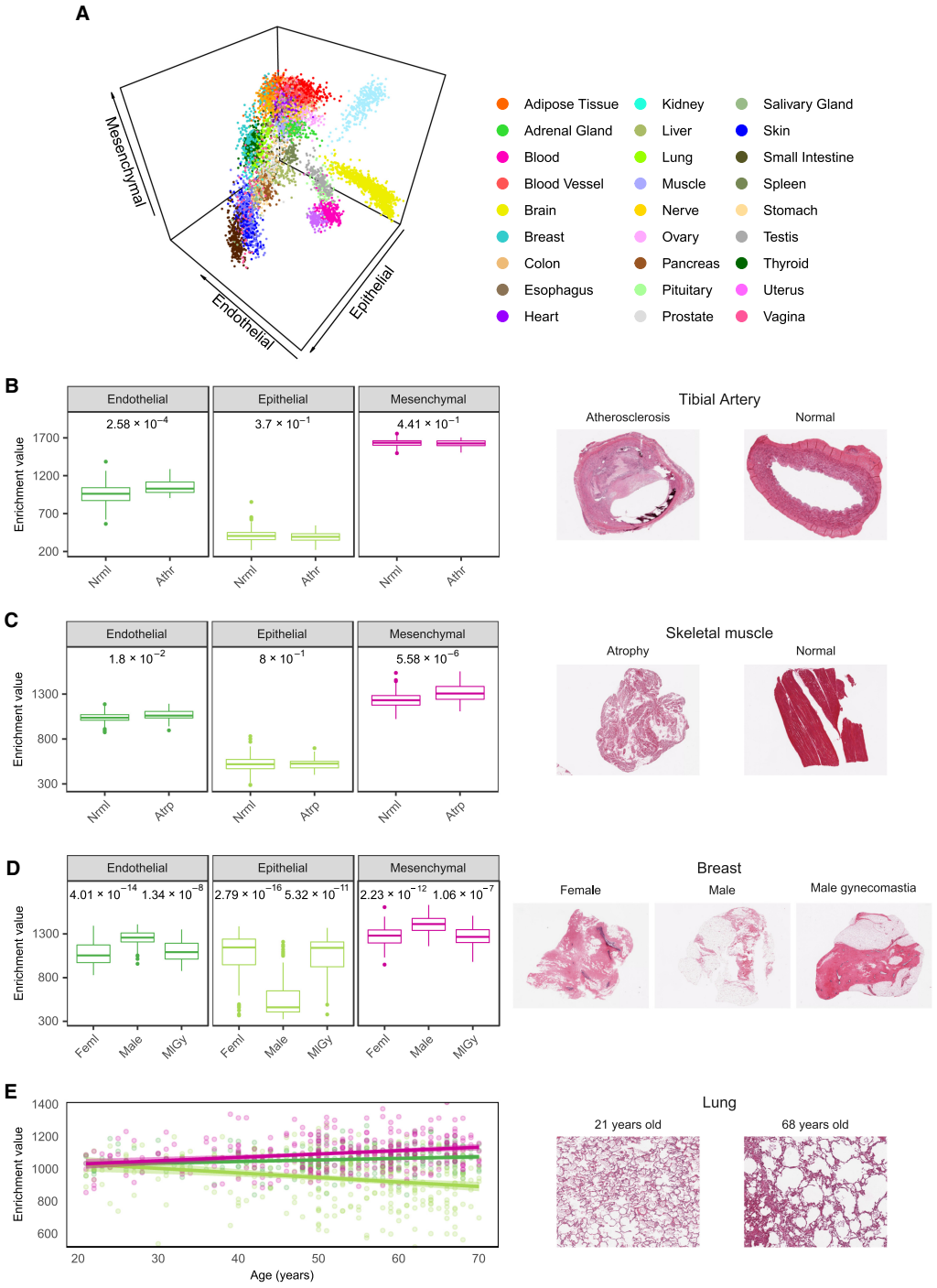


Figure 5. Alterations of the contributions of the major cell types to tissues in histological phenotypes. (*A*) GTEx samples represented in a 3D space in which the axes are the enrichments of endothelial, epithelial, and mesenchymal cells. (*B,C*) Differences in xCell enrichments of major cell types (Mann–Whitney *U* test, adjusted *P*-values as FDR) between affected and normal states. Histological images of affected and normal tissues are displayed (see text for details): (Athr) atherosclerosis ($n = 31$); (Atrp) atrophy ($n = 34$); (Nrml) normal ($n = 285$ and $n = 388$, respectively). (*D*) Major cell type xCell enrichments in female breast samples (Feml, $n = 85$), and male breast samples with (MlGy, $n = 36$) or without gynecomastia (Male, $n = 85$). Only significant FDR (≤ 0.05) are shown, all of them being between female and male without gynecomastia (*left*, FDR) and between male without gynecomastia and male with gynecomastia (*right*, FDR). (*E*) Changes in major cell type xCell enrichments in lung samples with age. Pearson's *r* and adjusted *P*-values as FDR: endothelial $r = 0.17$ and FDR $= 3.2 \times 10^{-3}$; epithelial $r = -0.23$ and FDR $= 6 \times 10^{-5}$; mesenchymal $r = 0.25$ and FDR $= 2.4 \times 10^{-5}$.

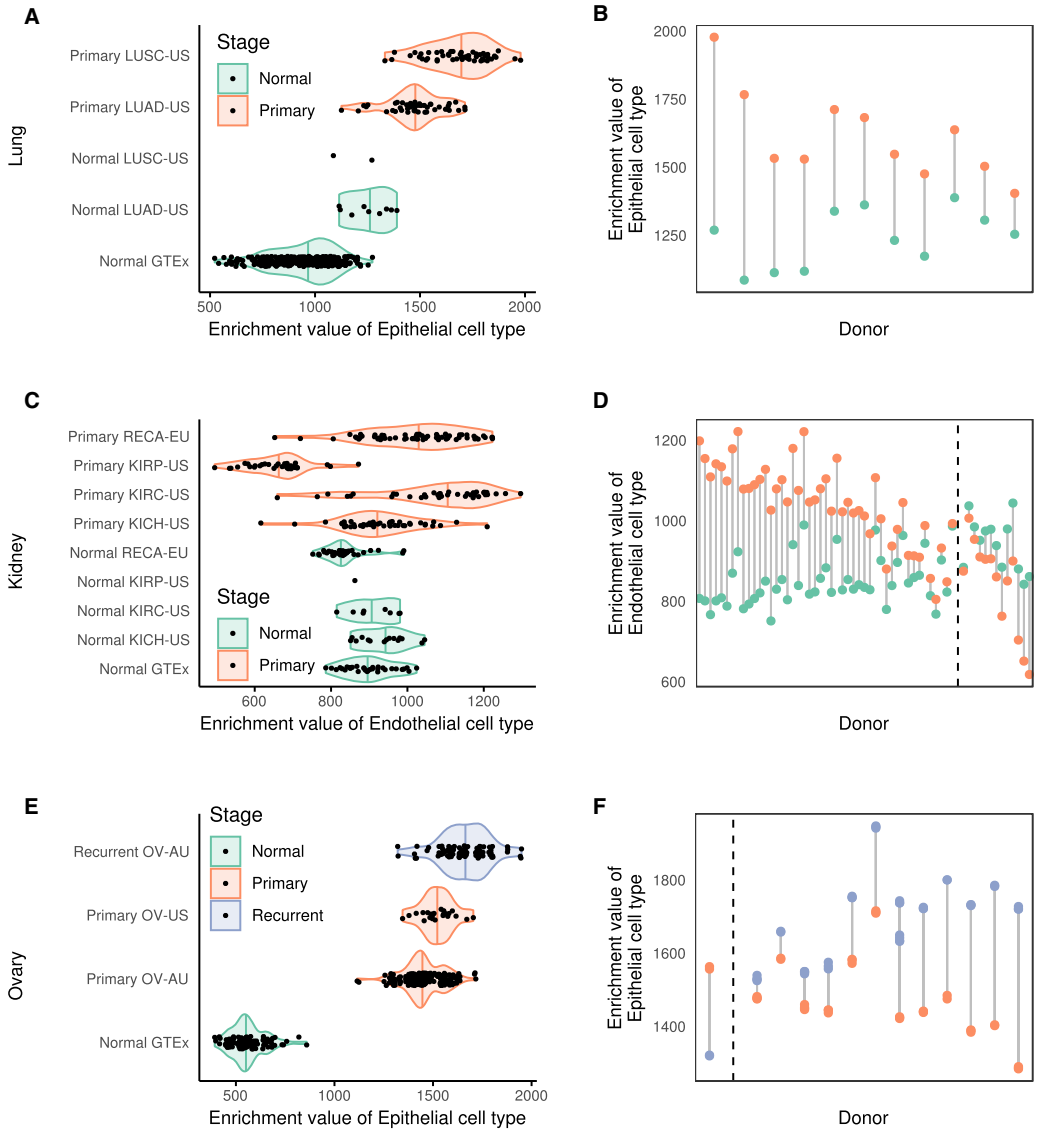


Figure 6. Alterations of the contributions of the major cell types to tissues in cancer. (*A*) xCell enrichments in epithelial cells in lung cancers and matched normal controls from the PCAWG project separated by cancer project: (LUAD-US) lung adenocarcinoma, TCGA, USA; (LUSC-US) lung squamous cell carcinoma, TCGA, USA. (*B*) Enrichment in matched normal and cancer lung samples by donor, pooled across the cancer projects. The *P*-value for the Mann–Whitney *U* test for the differences in epithelial contribution between normal and cancer samples in the LUAD-US project is: 8.1×10^{-6}. (*C*) xCell enrichment in endothelial cells in kidney cancers and matched normal controls from the PCAWG project separated by cancer project. (RECA-EU) renal cell cancer, France, EU; (KIRP-US) kidney renal papillary cell carcinoma, TCGA, USA; (KIRC-US) kidney renal clear cell carcinoma, TCGA, USA; (KICH-US) kidney chromophobe, TCGA, USA. (*D*) xCell enrichments in matched normal and cancer kidney samples by donor. The adjusted *P*-values for the Mann–Whitney *U* test for the differences in endothelial contribution between normal and cancer samples in the RECA-EU, KIRC-US, KICH-US projects are respectively 3.8×10^{-12}, 0.0024, and 0.65. (*E*,*F*) xCell enrichments in epithelial cells in ovarian cancers from the PCAWG project separated by cancer project (*E*) or by donor for matched primary and recurrent samples (*F*): (OV-AU) ovarian cancer, Austria; (OV-US) ovarian serous cystadenocarcinoma, TCGA, USA. The *P*-value for the Mann–Whitney *U* test for the differences in endothelial contribution between primary and recurrent samples in the OV-AU project is 3.6×10^{-27}. The donors in *B*, *D*, and *F* are sorted based on the difference between the enrichments. The dashed lines in *D* and *F* separate the matched samples in which the enrichment of endothelial (epithelial) cells is larger in the cancer sample from those in which it is larger in the normal sample.

intermediate levels of biological organization, namely, cells and tissues (or organs), is largely unknown. The reason has been the lack of phenotypic data on cells and tissues with associated genomic, epigenomic, and transcriptomic data.

Very recently, however, mostly through advances both in single-cell sequencing and in digital imaging technologies, data have started to become available, which can be used to connect the molecular to the cellular level, and this, in turn, to the tissue level. In

this regard, the data collected here on the transcriptomics of human primary cells, and the links that we have established between these data and the phenotypic traits of organs constitute a unique resource, serving as an intermediate resolution of complexity between single-cell and whole-organ transcriptomics. This resource will contribute to the understanding of how the interplay between transcription and cellular composition shapes tissue histology and ultimately impacts human phenotypes. Our analyses suggest that a large fraction of human cells and cell types in tissues belong to a few major cell types, providing a high-level transcriptionally based hierarchical classification of human cells. Extending the variety of profiled cell types, achieving single-cell resolution, and integrating expression data with epigenetics data, as proposed in the HCA project (Regev et al. 2017), will enrich our understanding of the constitutive cell types in the human body and their functional relationship.

Methods

RNA isolation, library construction, and sequencing

For each cell type to be made into a library, we obtained cell pellets that were stored in RNAlater (Thermo Fisher Scientific) as catalog items from PromoCell (https://www.promocell.com) and ScienCell (https://www.sciencellonline.com/) (for a list of primary cells, see Supplemental Table S1). In short, the RNA was isolated from sorted cells based on cell morphology and cell surface markers. Each cell type was passaged to expand the cell numbers for 24–48 h (1-2 doublings) before total RNA extraction and shipping. Thus, this protocol represents a minimum of exposure to non-native conditions. The cell morphologies are checked at this time. Although it is clear that the molecular context (influence of external cytokins and neighboring cells) of these cells has changed, they cluster in a very similar fashion to profiles shown by single-cell isolates of the corresponding types. Thus, the limited passage has an unlikely effect on the gene expression program. We rely on the providers' standards for quality assurance. Quality sheets are available through the ENCODE portal (https://www.encodeproject.org/search/?type=Biosample&organism.scientific_name=Homo+sapiens&biosample_ontology.classification=primary+cell&lab.title=Thomas+Gingeras%2C+CSHL&source.title=PromoCell&award.rfa=ENCODE3). We ordered three vials per cell type per donor for a total of 3 million cells. The three vials were combined, and we isolated total RNA from them using the Ambion mirVana miRNA Isolation kit (AM1561). The rRNA was removed using the RiboZero Gold Protocol (RZG1224). The libraries are made using a homebrew "dUTP" protocol (Parkhomchuk et al. 2009), which generates stranded libraries. They were sequenced on the Illumina platform in mate-pair fashion and processed through the data processing pipeline at the ENCODE DCC. Additional information about each of these steps, metadata, and files can be found at https://www.encodeproject.org/.

RAMPAGE sample preparation

Isolation of RNA is described in the preceding section. The RAMPAGE protocol (Batut and Gingeras 2013) was used to make libraries. Each library was sequenced in mate-pair fashion on the Illumina platform. Detailed protocol and quality-control images and metrics on a per library basis can be found in the "Production Documents" appended to each RAMPAGE assay at the ENCODE Data Coordination Center (https://www.encodeproject.org/).

Small RNA isolation, library construction, and sequencing

Isolation of RNA is described in the preceding section. The Illumina TruSeq protocol was used to make libraries. Each library was sequenced in single end fashion on the Illumina platform. Detailed protocol and quality-control images and metrics on a per library basis can be found in the "Production Documents" appended to each Small RNA assay at the ENCODE Data Coordination Center (https://www.encodeproject.org/).

RNA-seq processing pipeline

Raw reads from the 106 RNA-seq libraries (for a list of ENCODE library IDs, see Supplemental Table S1; for submitted FASTQ files, see https://www.encodeproject.org/) were aligned with STAR v2.3.1z (Dobin et al. 2013) to the human genome assembly hg19. Reads mapping to more than 20 multiple positions were discarded. Read counts for all long genes annotated in GENCODE v19 (Harrow et al. 2012) were computed with RSEM 1.2.19 (expected read counts) (Li and Dewey 2011). Statistics on the number of reads and mapping are available on Supplemental Table S14. Furthermore, we verified using *liftOver* that the cell-type-specific genes are consistent between GRCh37/hg19 and GRCh38/hg38, with a successful conversion of 2855 of the 2871 genes.

For most of the analyses, we average expression values for a given pair of replicates and sometimes the two biological replicates are from donors of the opposite sex; therefore, we remove genes on Chromosome Y. The lack of an enrichment step for polyadenylated transcripts preserves the presence of some short biotype genes, which are still longer than 200 bp. Thus, we remove genes with at least one transcript annotated as short RNA in GENCODE. These genes are often of repetitive nature, which makes the quantification of their expression problematic; this is why we decided to remove them.

Read counts which are not reproducible between two replicates (npIDR > 0.1) (Djebali et al. 2012) are set to 0. The matrix of read counts after npIDR is provided as Supplemental Table S2. After filtering for reproducibility, read counts are normalized to a slightly modified version of RPKM (reads per kilobase of exon model per million mapped reads) (Mortazavi et al. 2008). Specifically, read counts were first normalized to counts per million (cpm), in which the library sizes are the trimmed mean of M values (TMM) (Robinson and Oshlack 2010) scaled sums of exonic reads, and then normalized by gene length. Finally, RPKM values from the two replicates were averaged, and genes with RPKM < 1 in all samples were discarded, resulting in 16,265 genes, including 13,990 protein coding, 1380 long noncoding RNAs, and 895 pseudogenes. Statistical analyses were performed with R version 3.6.1 (R Core Team 2019).

As the samples were prepared and sequenced in three known distinct batches (Supplemental Table S1), we used the *removeBatchEffect()* function from R limma package (Ritchie et al. 2015) to build a linear model with the batch information and the cell types on $\log_{10}$-transformed RPKM (with a pseudocount of 0.01), and we regressed out the batch variable.

Data access

All experimental protocols for the samples described here, and all data generated for this study, are publicly available on the ENCODE portal (https://www.encodeproject.org/). GTEx gene expression is available in the GTEx portal (https://www.gtexportal.org).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This project was supported by awards U54HG007004, U41HG007234, and R01MH101814 from the National Human Genome Research Institute of the National Institutes of Health, as well as from the Spanish Ministry of Economy and Competitiveness, Centro de Excelencia Severo Ochoa 2013–2017, SEV-2012-0208, Programa de Ayudas FPI del Ministerio de Economía y Competitividad BES-2012-055848 to A.B., and Ministerio de Educación, Cultura y Deporte, under the FPU programme (Formación de Profesorado Universitario) with predoctoral fellowship FPU15/03635 to M.M.A., as well as the support of the CERCA programme/Generalitat de Catalunya. D.G.M. is supported by a "la Caixa"-Severo Ochoa predoctoral fellowship LCF/BQ/SO15/52260001. We also acknowledge support from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement 294653. We thank Kristin Ardlie and Detlev Arendt for useful discussions. We acknowledge and thank the donors and their families for their generous gifts of organ donation for transplantation and tissue donations for the GTEx research study. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (https://commonfund.nih.gov/GTEx). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We acknowledge the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership. Figure 1A was created with https://biorender.com/. R.G. dedicates this work to the Catalan leaders and the people in jail and exile for defending freedom and democracy, without which science cannot flourish.

Author contributions: A.B., C.A.D., M.M.A., V.W., R.G., and T.R.G. conceived and designed the experiments and analyses. J.D., C.A.D., A.S., and C.D. performed the experiments. A.B., M.M.A., V.W., and D.G.M. analyzed the data. J.G., D.D.P., A.V., A.D., C.Z., D.G.M., F.R., and M.P.S. contributed ideas and statistical advice. A.B., M.M.A., V.W., R.G., and T.R.G. wrote the manuscript.

References

Addison O, Marcus RL, LaStayo PC, Ryan AS. 2014. Intermuscular fat: a review of the consequences and causes. *Int J Endocrinol* **2014:** 309570. doi:10.1155/2014/309570

Appell HJ. 1990. Muscular atrophy following immobilisation. *Sports Med* **10:** 42–58. doi:10.2165/00007256-199010010-00005

Aran D, Hu Z, Butte AJ. 2017. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* **18:** 220. doi:10.1186/s13059-017-1349-1

Aziz SA, Sznol J, Adeniran A, Colberg JW, Camp RL, Kluger HM. 2013. Vascularity of primary and metastatic renal cell carcinoma specimens. *J Transl Med* **11:** 15. doi:10.1186/1479-5876-11-15

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338:** 1587–1593. doi:10.1126/science.1230612

Batut P, Gingeras TR. 2013. RAMPAGE: promoter activity profiling by paired-end sequencing of 5′-complete cDNAs. *Curr Protoc Mol Biol* **104**. doi:10.1002/0471142727.mb25b11s104

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The cancer genome atlas pan-cancer analysis project. *Nat Genet* **45:** 1113–1120. doi:10.1038/ng.2764

Cuhaci N, Polat S, Evranos B, Ersoy R, Cakir B. 2014. Gynecomastia: clinical evaluation and management. *Indian J Endocrinol Metab* **18:** 150–158. doi:10.4103/2230-8210.129104

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489:** 101–108. doi:10.1038/nature11233

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Elpek GÖ. 2014. Cellular and molecular mechanisms in the pathogenesis of liver fibrosis: an update. *World J Gastroenterol* **20:** 7260–7276. doi:10.3748/wjg.v20.i23.7260

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al. 2020. Expanded encyclopedias of DNA elements in the human and mouse genomes. *Nature* (in press) doi:10.1038/s41586-020-2493-4

Eroschenko VP. 2013. *DiFiore's atlas of histology with functional correlations*. Lippincott Williams & Wilkins, Baltimore.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507:** 462–470. doi:10.1038/nature13182

Frontini A, Giordano A, Cinti S. 2012. Endothelial cells of adipose tissues: a niche of adipogenesis. *Cell Cycle* **11:** 2765–2766. doi:10.4161/cc.21255

Gonzalez-Porta M, Calvo M, Sammeth M, Guigo R. 2012. Estimation of alternative splicing variability in human populations. *Genome Res* **22:** 528–538. doi:10.1101/gr.121947.111

The GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550:** 204–213. doi:10.1038/nature24277

Guo JH, Huang Q, Studholme DJ, Wu CQ, Zhao Z. 2005. Transcriptomic analyses support the similarity of gene expression between brain and testis in human as well as mouse. *Cytogenet Genome Res* **111:** 107–109. doi:10.1159/000086378

Haque A, Engel J, Teichmann SA, Lönnberg T. 2017. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* **9:** 75. doi:10.1186/s13073-017-0467-4

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774. doi:10.1101/gr.135350.111

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12:** 323. doi:10.1186/1471-2105-12-323

Manini TM, Clark BC, Nalls MA, Goodpaster BH, Ploutz-Snyder LL, Harris TB. 2007. Reduced physical activity increases intermuscular adipose tissue in healthy young adults. *Am J Clin Nutr* **85:** 377–384. doi:10.1093/ajcn/85.2.377

McLaughlin F, Ludbrook VJ, Cox J, von Carlowitz I, Brown S, Randi AM. 2001. Combined genomic and antisense analysis reveals that the transcription factor Erg is implicated in endothelial cell differentiation. *Blood* **98:** 3332–3339. doi:10.1182/blood.V98.12.3332

Mescher AL. 2013. *Junqueira's basic histology: text and atlas*. McGraw-Hill Medical, New York.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628. doi:10.1038/nmeth.1226

Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37:** e123. doi:10.1093/nar/gkp596

Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, Tanzer A, Lagarde J, Zaleski C, See LH, et al. 2015. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat Commun* **6:** 5903. doi:10.1038/ncomms6903

Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, Zhu J, Ohler U. 2011. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet* **7:** e1001274. doi:10.1371/journal.pgen.1001274

R Core Team. 2019. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. 2017. The human cell atlas. *eLife* **6:** e27041. doi:10.7554/eLife.27041.001

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43:** e47. doi:10.1093/nar/gkv007

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11:** R25. doi:10.1186/gb-2010-11-3-r25

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* **23:** 777–788. doi:10.1101/gr.152140.112

Sokocevic D, Bonenfant NR, Wagner DE, Borg ZD, Lathrop MJ, Lam YW, Deng B, DeSarno MJ, Ashikaga T, Loi R, et al. 2013. The effect of age and emphysematous and fibrotic injury on the re-cellularization of de-cellularized lungs. *Biomaterials* **34:** 3256–3269. doi:10.1016/j.biomaterials.2013.01.028

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep* **3:** 2179–2190. doi:10.1016/j.celrep.2013.05.031

The Tabula Muris Consortium. 2018. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562:** 367–372. doi:10.1038/s41586-018-0590-4

Tatone C, Amicarelli F, Carbone MC, Monteleone P, Caserta D, Marci R, Artini PG, Piomboni P, Focarelli R. 2008. Cellular and molecular aspects of ovarian follicle ageing. *Hum Reprod Update* **14:** 131–142. doi:10.1093/humupd/dmm048

Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res* **25:** 1491–1498. doi:10.1101/gr.190595.115

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32:** 381–386. doi:10.1038/nbt.2859

Yang RY, Quan J, Sodaei R, Aguet F, Segrè AV, Allen JA, Lanz TA, Reinhart V, Crawford M, Hasson S, et al. 2018. A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. bioRxiv doi: 10.1101/311563

Yoh K, Prywes R. 2015. Pathway regulation of p63, a director of epithelial cell fate. *Front Endocrinol (Lausanne)* **6:** 51. doi:10.3389/fendo.2015.00051

Young B, O'Dowd G, Woodford P. 2013. *Wheater's functional histology*. Elsevier Health Sciences, New York.

Zhang HM, Chen H, Liu W, Liu H, Gong J, Wang H, Guo AY. 2012. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* **40:** D144–D149. doi:10.1093/nar/gkr965

Ziegler MA, Distasi MR, Bills RG, Miller SJ, Alloosh M, Murphy MP, Akingba AG, Sturek M, Dalsing MC, Unthank JL. 2010. Marvels, mysteries, and misconceptions of vascular compensation to peripheral artery occlusion. *Microcirculation* **17:** 3–20. doi:10.1111/j.1549-8719.2010.00008.x

Received March 10, 2020; accepted in revised form April 29, 2020.

A limited set of transcriptional programs define major cell types

Alessandra Breschi, Manuel Muñoz-Aguirre, Valentin Wucher, et al.

Genome Res. 2020 30: 1047-1059 originally published online July 29, 2020
Access the most recent version at doi:10.1101/gr.263186.120

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

The impact of sex on gene expression and its genetic regulation across human tissues

Mechanisms that have a role in the biology of human sex can translate to complex phenotypes and diseases, for which some can exhibit different behaviors in males and females. Although factors such as the sex chromosomes and hormones are typically attributed as sources of these effects, less is known about sex differences at the molecular level. To this end, in this work we aim to document sex differences in the human transcriptome and its regulation. We survey sex differences (defining males and females solely on the basis of their XY and XX sex chromosomes) in the transcriptome of 44 human tissues, and find that differential gene expression with respect to sex is ubiquitous across tissues but with small effect sizes, and tends to be largely tissue specific. Using computationally-inferred cell type enrichments from bulk RNA-seq gene expression (with different gene signatures than those used in Chapter 4), we characterize differences in cell type composition between males and females and how these can be linked to phenotypes associated to disease. We find that the sex differences in genetic regulation (sex-biased eQTLs) are much less common when compared to gene expression differences. Nevertheless, we find associations for gene-trait pairs that are driven by a single sex.

Summary of my key contributions

- Estimated surrogate variables from gene expression data in order to capture variation from unknown technical sources. Demonstrated that these surrogate variables are correlated with computationally-inferred cell type abundances, as well with other technical covariates although to a lesser degree.

- Performed differential gene expression between males and females (948 human individuals) separately in 44 tissues totalling 16,245 RNA-seq samples across 35,431 genes. With these results, we found that:

    1. Sex effects on gene expression exist across all tissues. A total of 13,294 genes (37 % of the transcriptome) exhibit differential expression in at least one tissue.

    2. Sex-differential expression effect sizes are small, especially for autosomal genes.

    3. Sex-biased genes are, in general, tissue-specific and tend to be expressed at most in small subsets of tissues. Only 30 genes have a consistent sex bias across all tissues, and most of these are known XCI escapees.

- Assessed the replicability of differentially expressed genes identified with GTEx RNA-seq by performing differential expression on independent datasets in four different tissues, observing moderate to strong replication. Containerized all the differential expression analyses in a reproducible Docker image.

- Illustrated tissue similarities with hierarchical clustering (with uncertainty assessment via multiscale bootstrap resampling) based on gene expression profiles and differential expression (meta-analysis) effect sizes. Identified driver genes for each cluster, both in the clustering based on gene expression and on sex-differential expression effect sizes. With these, we observed that:

    – Hierarchical clustering between both data types is highly concordant.

  - Tissue specificity of sex-differentiated effects is not driven by tissue-specific gene expression (i.e. the clusters, although similar, are not primarily weighted by the same genes).

- Developed cross-validated gradient boosted tree classifier models to predict sex in for each of the 44 tissues, performing parameter search with bayesian optimization and Gaussian processes. Computed Shapley values for each model as a proxy to explain class predictions at the sample level. With these, we found that:

  1. Prediction with X-linked genes is accurate, with high sensitivity and specificity in all tissues. The most predictive X-linked genes are known X-Chromosome Inactivation (XCI) escapees.
  2. 40 X-linked genes that are predictive of sex have not been previously described as X-Chromosome Inactivation (XCI) escapees, suggesting their potential escapism, which needs to be validated experimentally.
  3. Sex prediction with autosomal-only genes is less accurate, less specific and required more genes. The exceptions are breast and muscle tissues, where autosomal genes predicted sex with specificity and sensitivity $\geq 90\%$.
  4. Genes among the top predictive autosomal ones have been shown to have roles in hormone regulation and sex-differentiated traits.

- Identified sets of transcription factor (TF) binding sites with large enrichment differences between female- and male- biased genes. Although some of these TFs have known roles in sex differences, for example, in body growth rates and liver gene expression, the rest are mostly uncharacterized.

- Analyzed sex differences in computationally-inferred enrichments for seven cell types. Found that significant differences exist in a few tissues.

- Found six pathological phenotypes with altered cell type enrichments in males and females, using the histological phenotypes derived from free-text pathology reports in Chapter 4.

- Performed PCA of gene expression in breast tissue using the set of autosomal predictive genes derived from the classifer model mentioned above. Found that this set of genes, besides allowing to discriminate between males and females, also identifies a separation between male individuals affected by gynecomastia, and non-affected males. By examining the loadings, we characterize a set of genes that show association with gynecomastic males, an example is the *TRH*, which is known to play a role in the mediation of protein levels which are altered in gynecomastia.

Note:

Due to licensing restrictions, here follows the accepted version of the manuscript. Please refer to https://doi.org/10.1126/science.aba3066 for the copyedited (published) version of the manuscript.

Accepted version of the manuscript: Oliva, M., Muñoz-Aguirre, et al. (2020). The impact of sex on gene expression across human tissues. Science, 369(6509), eaba3066. https://doi.org/10.1126/science.aba3066. Reprinted with permission from AAAS. License number: 4952010594208

Title: The impact of sex on gene expression across human tissues

Authors: Meritxell Oliva[1,2,3*†], Manuel Muñoz-Aguirre[4,5†], Sarah Kim-Hellmuth[6,7,8†], Valentin Wucher[4], Ariel D.H. Gewirtz[9], Daniel J. Cotter[10], Princy Parsana[11], Silva Kasela[7,8], Brunilda Balliu[12], Ana Viñuela[13], Stephane E. Castel[7,8], Pejman Mohammadi[14], François Aguet[15], Yuxin Zou[16], Ekaterina A. Khramtsova[1,17], Andrew D. Skol[1,2,18,19], Diego Garrido-Martín[4], Ferran Reverter[20], Andrew Brown[21], Patrick Evans[22], Eric R. Gamazon[22,23], Anthony Payne[24], Rodrigo Bonazzola[1], Alvaro N. Barbeira[1], Andrew R. Hamel[15,25], Angel Martinez-Perez[26], José Manuel Soria[26], GTEx Consortium, Brandon L. Pierce[3], Matthew Stephens[16,27], Eleazar Eskin[28], Emmanouil T. Dermitzakis[13], Ayellet V. Segrè[15,25], Hae Kyung Im[1], Barbara E. Engelhardt[29], Kristin G. Ardlie[15], Stephen B. Montgomery[10,30], Alexis J. Battle[11,31], Tuuli Lappalainen[7,8], Roderic Guigó[4,32], and Barbara E. Stranger[1,2,18, 33*]
† Contributed equally to this work
* Corresponding author. Email: meritxellop@uchicago.edu, barbara.stranger@northwestern.edu

Affiliations:
[1] Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA
[2] Institute for Genomics and Systems Biology, The University of Chicago, Chicago, IL, USA
[3] Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA
[4] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Catalonia, Spain
[5] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain
[6] Statistical Genetics, Max Planck Institute of Psychiatry, Munich, Germany
[7] New York Genome Center, New York, NY, USA
[8] Department of Systems Biology, Columbia University, New York, NY, USA
[9] Department of Computer Science, Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, USA
[10] Department of Genetics, Stanford University, Stanford, CA, USA
[11] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
[12] Department of Computational Medicine, University of California, Los Angeles, CA, USA
[13] Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland
[14] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, Scripps Research Translational Institute, La Jolla, CA, USA
[15] The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA.
[16] Department of Statistics, The University of Chicago, Chicago, IL, USA
[17] Computational Sciences, Janssen Pharmaceuticals, Spring House, PA, USA
[18] Center for Translational Data Science, The University of Chicago, Chicago, IL, USA
[19] Department of Pathology and Laboratory Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, USA
[20] Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona. Barcelona, Spain

[21] University of Dundee, Dundee, Scotland, UK
[22] Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[23] Clare Hall, University of Cambridge, Cambridge, England, UK
[24] Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK.
[25] Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA
[26] Genomics of Complex Diseases Group. Research Institute Hospital de la Sant Creu i Sant Pau. IIB Sant Pau. Barcelona, Spain.
[27] Department of Human Genetics, The University of Chicago, Chicago, IL, USA
[28] Departments of Computational Medicine, Computer Science, and Human Genetics, University of California, Los Angeles, CA, USA
[29] Department of Computer Science, Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, USA
[30] Department of Pathology, Stanford University, Stanford, CA, USA
[31] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[32] Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
[33] Center for Genetic Medicine, Department of Pharmacology, Northwestern University, Chicago, IL USA

Abstract: Many complex human phenotypes exhibit sex-differentiated characteristics. However, the underlying molecular mechanisms of these differences remain largely unknown. We generated a catalog of sex differences in gene expression and in the genetic regulation of gene expression across 44 human tissue sources surveyed by GTEx (v8 release). We demonstrate that sex influences gene expression levels and cellular composition of tissue samples across the human body. A total of 37% of all genes exhibit sex-biased expression in at least one tissue. We identify *cis*-eQTLs with sex-differentiated effects and characterize their cellular origin. By integrating sex-biased eQTLs with genome-wide association study data, we identify 58 gene-trait associations that are driven by genetic regulation of gene expression in a single sex. Together we provide an extensive characterization of sex differences in the human transcriptome and its genetic regulation.

One Sentence Summary: Sex differences in the human transcriptome are widespread, highly tissue-specific, and contribute to our understanding of sex-differentiated biology and complex traits.

Main Text

Many complex human phenotypes, such as anthropometric traits like waist-to-hip ratio, exhibit sex-differentiated distributions; disease features, such as prevalence, progression, age of onset, and response to treatment, often differ by sex (*1*–*5*). These sex differences have been previously attributed to hormones, sex chromosomes, differences in behavior and environmental exposures, as well as various genetic models (*6*), but the mechanisms and underlying biology of the sex differences remain largely unknown. The Genotype-Tissue Expression (GTEx) project (*7*) provides an opportunity to investigate the prevalence and genetic mechanisms of sex differences in transcriptomes and to identify how sex and genetics interact to influence complex traits and disease. The analyses presented here characterize sex differences in a relatively large population sample, including many tissues that generally lack characterization. As the causative tissue is unknown for many diseases and disorders, analysis of this diverse tissue set can serve as a powerful resource for investigations into the basis of sex-differentiated phenotypes.

We present an extensive characterization of sex differences in the human transcriptome across 44 tissue sources of the GTEx project (v8 data release, (*8*)) from 838 individuals (557 males, 281 females), comprising a large collection of multi-tissue bulk gene expression and genotype data (Fig. 1, (*9*)). Here we quantify and characterize sex differences in gene expression levels (sex-biased gene expression) and *cis* sex-biased expression quantitative trait loci (sb-eQTLs). By incorporating the results of these sex-aware analyses of GTEx data with gene features and transcription factor binding annotation, we describe tissue-specific and tissue-non-specific drivers and mechanisms contributing to sex differences in the human transcriptome and eQTLs. By integrating data from genome-wide association studies (GWAS), we report multiple sex-differentiated genetic effects on the transcriptome that colocalize with complex trait associations, highlighting the power of characterizing sex bias in GTEx samples for the mechanistic interpretation of GWAS signals.

Sex effects on gene expression are ubiquitous but small

Using GTEx v8 data (Table S1), we quantified sex-biased gene expression in each of the 44 tissue sources for all genes expressed in at least one tissue. We consider a total of 35,431 X-linked and autosomal genes, including protein-coding, lincRNA, and other less characterized gene types such as transcribed pseudogenes (*9*). For each tissue, we first fit a linear model that accounts for known sample and donor characteristics, as well as surrogate variables that capture hidden technical or biological factors of expression variability, including tissue cell-type composition (Fig. S1, A, B and C). Consequently, we are able to identify sex-biased gene expression that does not derive from sex differences in cell-type abundances. We next modeled sex-bias effects across tissues. We discovered a total of 13,294 differentially expressed genes (sex-biased genes; local false sign rate (LFSR) ≤ 0.05), with 473 to 4,558 genes discovered per tissue, representing 1.3% to 12.9% of all tested genes, respectively (Figs. 2A, S1, D to F, Table S2). Previous studies have reported widespread sex-biased gene expression (*10*–*12*), and described breast as the most sex-differentiated tissue (*10*, *11*, *13*). However, we did not observe this in the present study after controlling for sex differences in tissue cell-type composition (Fig. S1A). We next assessed replication of sex-biased genes in independent gene expression datasets for four tissues (brain cerebellum, brain cortex, heart left ventricle and lymphocytes, Table S2). We observed moderate to strong replication (average π_1=0.62, average effect size Spearman's rho = 0.78). In total, 37.5%

(13,294/35,431) of the human transcriptome was differentially expressed in at least one tissue. Of these, 531 genes (4%) were X-linked and 12,763 genes (96%) were autosomal, representing 47% and 37% of all tested X-linked and autosomal genes, respectively. Although abundant, sex effects are mostly small (Fig. S2A), particularly for autosomal genes ((*9*), Fig. S2B). X-linked genes with higher expression in females (female-biased genes) exhibited larger sex effects (median fold change (FC) = 1.13) than either X-linked genes with higher expression in males (male-biased genes; median FC = 1.08), or autosomal sex-biased genes (median FC_M and FC_F = 1.04; Fig. S2B), potentially due to escape from X-chromosome inactivation (XCI) (*14*). The number of sex-biased genes and the effect sizes were not dominated by either sex (Fig. S2C).

Sex-biased gene expression is largely tissue-specific

Sex-biased genes exhibited a skewed pattern of tissue sharing - they were likely to be differentially expressed in only a small subset of tissues (Fig. 2B), as previously reported (*10–13*). Of 13,294 total sex-biased genes, 2,416 were (18.2%) differentially expressed in only a single tissue (Fig. 2B), suggesting tissue-dependent regulation. Only 30 genes (0.23%; 22 of which are known constitutive XCI escapees, Table S3) exhibited consistent sex bias across all 44 tissue sources (Fig. 2B). This tissue-specificity did not simply reflect patterns of gene expression across tissues; sex-biased genes tended to be ubiquitously expressed across tissues, while sex-biased expression was limited to one or a few tissues ((*9*), Figs. 2C, S2D). The majority (8,241/10,878 genes, 76%) of genes with sex bias in ≥2 tissues exhibit consistent effect direction across tissues, especially for X-linked genes (Fig. S2E). Notably, whole blood and cell lines, the most widely studied biospecimen types, were not representative of sex-biased expression across tissues; sex-biased genes in whole blood comprised only 12.9% (1,710/13,294) of all sex-biased genes. Although hierarchical clustering of tissues based on gene expression and sex-biased expression are highly concordant (Cophenetic correlation = 0.75; (*9*), Fig. 2C, Fig. S3, A, B and C), the intersection between the cluster-defining gene sets (Table S4) is less than expected by chance ($P < 2.2e\text{-}16$, Hypergeometric test). For example, both expression and sex-biased expression supported a cluster of brain subregions that is clearly differentiated from other tissues (Figs. 2C, S3, B and D). However, the cluster based on sex-biased expression was driven by 194 genes, while the transcriptome-based brain cluster was driven by 982 genes, from which only six were common with those defining the sex-based brain cluster. Among drivers of the sex-based liver cluster, we identified CYP450 genes - *CYP1A2*, *CYP3A7*, *CYP3A4* - as previously reported (*15*); but we also found genes less well-characterized for sex bias, such as *PZP*, *H19* and *VWCE* which were previously shown to be sex-differentially expressed due to liver-specific sex differences in DNA methylation (*16*). These results suggest that the tissue-specificity of sex-biased expression is not driven primarily by tissue-specific gene expression.

X-linked female-biased genes accurately predict sex and suggest tissue-specific candidates for escape from X-chromosome inactivation

We accurately predicted sex from gene expression, as previously explored (*17*), using X-linked genes ((*9*), Fig. S4, A to D) with gradient boosted trees. Although the most predictive X-linked genes (Fig. S4E) are those known to escape XCI, we identified 40 X-linked female-biased genes predictive of sex (within the top tertile with respect to their Shapley values) not previously described as XCI escapees (Table S3). While these results suggest further evaluation of these genes as potential XCI escapees, we did not directly test escape from XCI, and female-biased expression

of X-linked genes may originate from other mechanisms. Sex prediction from autosomal genes was less accurate (mean accuracy = 84%), less specific (mean specificity = 56% and sensitivity = 96%; Fig. S4D), and required more genes (Fig. S4F) than prediction based on X-linked genes. However, in two tissues - breast and muscle - autosomal genes predicted sex with specificity ≥ 90% and sensitivity ≥ 98% (Fig. S4G).

Sex-biased genes exhibit non-random and tissue-specific genomic distribution

Except for the enrichment of female-biased genes on the X chromosome, little is known about the genome-wide distribution of sex-biased genes. We applied a positional gene enrichment analysis method (PGE) (*18*) separately for male- and female-biased genes (LFSR ≤ 0.05) from each tissue ((*9*), Fig. S5A). We discovered clustering of a total of 1,559 sex-biased genes in 134 autosomal and five X-linked regions (hypergeometric test P ≤ 0.001; Fig. 3A, left; Table S5). On the X chromosome, pseudoautosomal (PAR) region PAR1 and the remainder of the X chromosome short arm *p* were enriched for male-biased and female-biased genes, respectively (Fig. 3A, right), as previously reported (*14*). Female-biased gene enrichment was stronger (Spearman's rho=0.51, P = 1.63e-15) in the younger strata of arm *p* (Fig. S5B), likely driven by escape from XCI (*14*, *19*). Although enriched X chromosome regions spanned ~126Mb, only 25% of subregions were enriched in at least two thirds of the tissues. Among autosomal sex-biased genes, we observed a cluster of male-biased genes on chromosome 20 that was identified in 70% (30/44) of tissues (Fig. S5C), but the majority of the 134 autosomal enriched regions were tissue-specific, identified on average in ~7% (3/44) of tissues (Fig. S5D, Table S5). These results are compatible with tissue-variable escape from XCI (*14*, *20*) and with tissue-specific topologically associating domains, possibly mediated by hormones (*21*). Further investigation is warranted to corroborate these and other hypotheses, as observed patterns may originate from a variety of mechanisms.

Promoters of sex-biased genes are enriched for hormone-related and other transcription factor binding sites

We hypothesized that transcription factor (TF) activity might drive observed patterns of differential expression, as sex-biased gene regulation by TFs has been recently reported (*13*), and TFs contribute to evolutionary changes in sex bias (*12*). We tested for enrichment of TF binding sites (TFBS) of 231 TFs previously identified through ChIP-seq (*22*) in promoter regions (i.e., 2kb upstream of the transcription start site) of male- and female-biased genes ((*9*), Fig. S5E). We discovered enrichment for TFBS of a total of 92 TFs (Fig. S5F), two of which were X-linked (*AR*, *ELK1*). TFBS for 54 TFs were enriched among female-biased genes, and 60 TFs among male-biased genes, with 22 TFs enriched among both sets of genes (Table S6). The 92 TFs include i) known hormone-related TFs estrogen (*ESR1*), androgen (AR) and glucocorticoid (*NR3C1*) receptors, ii) ten TFs that colocalize with steroid receptors, and iii) TFs with a non-reported or less-characterized hormone association, including *SP1*, *E2F6*, *NRF1*, *KLF9*, and *SP2*, the top five TFs with consistent TFBS enrichment across tissues (*9*).

The strongest difference between male- versus female-biased enrichment profiles was observed for TFBS of *SP2*, *SP4*, *NFYB*, *TWIST1*, *STAT5B* (female-biased) and *HNF4G*, *NFKB1*, *E2F6*, *HNF4A*, *ETS1* (male-biased), respectively, which were detected across most tissues (Fig. 3B, Table S6). In contrast, we observed tissue-specificity for enrichment of TFBSs of several TFs, e.g., *RFX2*, *ETV4* for brain and breast tissues, respectively (Fig. 3B, Fig. S5F). While *STAT5B* and *HNF4A* play known roles in sex differences in body growth rates and liver gene expression (*15*),

less is known about their roles and sex biases across all tissues. The effect of sex on most of the remaining TFs is uncharacterized. Together, these results suggest that hormone-related TFs regulate sex-biased expression as expected, but also indicate that additional TFs play a role in sex-biased gene expression, in some cases in a tissue-specific manner (Table S6). Notably, TFBS enrichment is not driven by sex-biased expression of the TFs themselves (*9*), consistent with the observation that sex-biased TF targeting of genes is independent of sex-biased gene expression (*13*). However, this scenario cannot be discarded if such differences occur at an earlier developmental time point and translate into a more constitutive sex-biased TF binding profile (*23*). Alternatively, other mechanisms involving TFs could be causal drivers, e.g., post-translational modifications as reported in mice (*24*).

Sex-biased genes are involved in a highly diverse set of biological functions and suggest sex-specific deposition of epigenetic marks

To gain insight into cellular functions affected by sex-biased genes, we performed gene set enrichment analysis (GSEA) in each tissue, considering the direction of the sex effect ((*9*), Fig. S6A, Table S7 and S8). To identify gene sets that are enriched across multiple tissues, we performed a meta-analysis using Fisher's combined probability test and identified 2,134 enriched gene sets (FDR $\leq$ 0.05; Table S9). We applied a community detection approach to identify common features across enriched gene sets and defined 36 clusters (Table S9). Among the top scoring clusters ((*9*)), we identified enrichment of genes in pathways involved in drug and hormone response, epigenetic marks, embryonic development and tissue morphogenesis, fertilization, sexual reproduction and spermatogenesis, fat metabolism, cancer, immune response, and other functions (Fig. 3C, Table S9). The top scoring cluster corresponds to targets of PRC2 and trimethylation of histone H3 at lysine 27 (H3K27me3), which is predominantly driven by female-biased genes; a pattern also reported for other epigenetic modifications (*13*). This complex induces gene silencing and is involved in XCI (*25*). Sex-specific deposition of H3K27me3 marks have been previously reported, resulting in sex-biased gene expression in mammalian placenta (*26*) and adult liver (*27*). These differences have been hypothesized to be regulated by sex differences in secretion of placental glycosyltransferase OGT and pituitary growth hormone. The observed association of H3K27me3 with sex biased expression in the tissues of this study (Table S9) has not been previously reported. We also identified clusters related to drug metabolism that include CYP450 genes. While sex-biased expression of CYP450 has been reported in liver (*15*) and linked to sex-differentiated growth hormone profiles, we observed sex-biased expression in additional tissues (Fig. S6B). Sex-biased expression was also identified for clusters related to gonad tissue functions, e.g., meiotic synapsis, which comprise genes expressed largely in testis (Fig. S6B). It is possible that part of the cross-tissue sex-biased expression patterns observed in adult tissues are derived from gamete formation and embryogenesis (*28*). Together, these results indicate that sex-biased genes are involved in a wide range of biological functions and pathways, many of which have not been previously associated with sex differences.

Sex and disease influence tissue cellular composition

The GTEx tissue samples are mixtures of heterogeneous cell types, with variation among individuals and tissues (*29*). In whole blood, cell-type composition differs between sexes (*30*, *31*), but little is known about sex differences in composition of other tissues. Using a t-test we examined each GTEx tissue for sex differences in cellular composition using estimated abundances of seven

cell types (*29*) (*9*). We discovered significant (FDR ≤ 0.05) differences for four cell types - keratinocytes, neutrophils, adipocytes and epithelial cells - in three tissues (Fig. S7A, Table S10). We hypothesize that additional cell types uncharacterized in this study may influence cell-type composition of GTEx tissues, particularly of immune cells, as marked sex differences in immune cell abundances have been reported (*30*, *32*). To investigate cellular abundances in disease, we utilized histological annotations from pathology review of GTEx tissue samples (*9*). We discovered six pathological phenotypes with altered cell-type composition (Fig. S7 B-E, Table S10). Together these results suggest that sex is correlated with tissue cellular composition, and that disease may alter cellular abundances in a sex-differentiated manner or in sex-specific pathologies.

Sex differences in the genetic regulation of gene expression are highly tissue-specific and less common than sex effects on gene expression

Sex-differentiated human phenotypes and disease characteristics may derive, in part, from sex-differentiated genetic effects (*6*, *33–36*), some of which may impact gene expression. For each of 491,694 conditionally independent *cis*-eQTLs identified in the sex-combined *cis*-eQTL analysis of the GTEx v8 project (*8*), we performed sex-biased *cis*-eQTL (sb-eQTL) analysis in each of 44 tissues present in both sexes (Fig. 1). We used a linear regression model including genotype, sex, and covariates, and tested for significance of a genotype-by-sex (G×Sex) interaction on expression (*9*). Notably, this approach captures both G×Sex interactions that derive from sex and from sex-correlated factors, including cell-type abundances or environmental factors. While the contribution of cell-type heterogeneity to sb-eQTLs is currently unknown, we observed sex differences in tissue cell-type composition (Fig S7A), which may impact sb-eQTL discovery. Hence, we characterized the impact of cell-type-specific eQTLs on sb-eQTLs (see below). We discovered a total of 369 sb-eQTLs, corresponding to 366 genes (sb-eGenes) (FDR ≤ 0.25, Table S11). The majority of sb-eQTLs were identified in breast (261 sb-eQTLs), but also in muscle (36 sb-eQTLs), skin (18 sb-eQTLs), and adipose tissues (14 sb-eQTLs; Fig. 4A, Fig. S8, A and B). Overall, sb-eQTLs showed strong evidence for tissue specificity (*9*); only one sb-eQTL was significant in two tissues (Table S11), and only 21% displayed patterns suggestive of tissue-sharing even at a lenient significance threshold ($P_{G\times Sex} \leq 0.01$). Only 36 (14%) sb-eGenes exhibited sex-biased expression in the discovery tissue (*MASH* LFSR ≤ 0.05; Table S12), similarly to recent observations (*37*). (*37*) different functional mechanisms contributing to each sex bias type.

To provide additional support for the sb-eQTLs, we used two approaches to assess differential allele-specific expression (ASE) between sexes: allelic fold-change (ASE aFC) (*38*) and Environment-ASE through Generalized LinEar modeling (EAGLE) (*39*) (*9*). Allele-specific expression can arise due to *cis*-regulatory genetic effects in heterozygous individuals. Differential ASE, therefore, indicates condition-specific *cis* effects (*39*) including sex-specificity. We observed that, despite limited power when restricted to heterozygous individuals and differences in methodology, both approaches support that a portion of the detected sb-eQTLs correspond to sex differences in ASE (Fig. S8C): sb-eQTLs were enriched for sex-biased ASE aFC (All tissues: $\pi_1 = 0.36$, breast: $\pi_1 = 0.41$, Fig. S8, D and E) and for EAGLE associations ($\pi_1 = 0.13$, empirical test, P ≤ 0.001). Of the 243 and 163 sb-eQTLs tested by ASE aFC and EAGLE methods, respectively, 65 (26.7%) were supported by ASE aFC (Wilcoxon P ≤ 0.05) (Fig. S7, F and G), 29 (17.8%) were supported by significant EAGLE associations, and 16 sb-eQTLs (10.4% of the 154 sb-eQTLs tested by both methods) were supported by both methods (Table S11).

We were limited in our ability to replicate sb-eQTLs because the majority of sb-eQTLs were discovered in breast, and matching, well-powered datasets do not exist. We performed internal validation, splitting GTEx breast samples into discovery and validation cohorts, and observed moderate replication (mean π_1 = 0.28; (*9*), Fig. S8H). We next assessed sb-eQTL replication (considering sb-eQTLs from breast, whole blood, and all tissues) in independent larger (~900 subjects) whole blood eQTL datasets, including DGN (*40*) and GAIT2 (*41*) cohorts ((*9*), Table S13). We observed weak replication (π_1 = 0 - 0.12, depending on sb-eQTL set and replication cohort). Poor replication of sb-eQTLs has been previously reported (*40*, *42*, *43*) and has been, in part, attributed to low power (*44*), but also to methodological and study design differences.

For each sb-eGene, we also performed sex-stratified *cis*-eQTL analysis for each tissue, downsampling males to match the female sample size (*9*). We observed strong correlation (Spearman's rank correlation rho = 0.78, P ≤ 2.2e-16) between male and female *cis*-eQTL effect sizes. For 58% of sb-eQTLs, sex-stratified *cis*-eQTL analysis revealed associations in both sexes with concordant allelic effect, but different effect sizes. For example, rs117380715-*ADRA1A* in adipose subcutaneous tissue showed a stronger effect in females than in males (β_F = -0.78, P_F = 4.64e-18, β_M = -0.47, P_M = 3.98e-10) (Fig. 4B, Fig. S8I). For the remainder of the sb-eQTLs, a *cis*-eQTL was detected exclusively in either females (70, 19%) or males (84, 23%). For example, we identified a female-specific *cis*-eQTL for rs8942-*C4BPB* in breast (β_F = 0.40, P_F = 2.68e-07, β_M = -0.02, P_M = 8.90e-01) (Fig. 4B, Fig. S8I). *C4BPB* encodes the beta unit of the C4b-binding protein, and controls activation of the complement cascade (*45*). We also identified a male-specific *cis*-eQTL for rs2273535-*AURKA* in skeletal muscle (β_M = 0.47, β_F = 0.01), described in (*8*). *AURKA*, Aurora kinase A, is a member of the serine/threonine kinase family involved in mitotic chromosomal segregation, muscle differentiation (*46*), and a known risk factor for several cancers (*47*). These results demonstrate that sex-biased genetic effects on gene expression exist for a small proportion of previously identified *cis*-eQTLs, and some sb-eQTLs affect genes implicated in human phenotypes.

Sex differences in genetic regulation of gene expression are partially mediated by cell-type-specific eQTLs

Given that the G×Sex interaction term of our eQTL model captures interactions that derive from sex as well as interactions with sex-correlated factors, we next characterized the fraction of sex-biased eQTLs that are driven by cell-type-specific eQTLs (Fig. S9A). We focused on breast, the tissue with the most sb-eQTLs and the largest sex differences in cellular composition (Figs. S7A, S8B). We tested 261 breast sb-eQTLs for enrichment of cell-type interacting *cis*-eQTLs (ieQTL) (*29*) (*9*). These ieQTLs correspond to *cis*-eQTLs where the effect varies depending on estimated cell-type abundances (*29*). Breast sb-eQTLs were strongly enriched (π_1 = 0.66 and 0.89) for ieQTL signal corresponding to adipocytes and epithelial cells (Fig. S9B). After including an interaction term for genotype-by-epithelial cell abundance estimates in the sb-eQTL model, 58% of breast sb-eQTLs (152/261) remained significant, while for 42% of sb-eQTLs (109/261), the genotype-by-sex effect was strongly attenuated (Fig. S9C, Table S14). For example, the strongest breast sb-eQTL, rs2289149-*LINC00920* (P = 4.83E-11) was not significant after incorporating the genotype-by-epithelial cell abundance estimates in the model ($\beta_{G\times Sex}$ = 0.187, CI(95%) = [-0.004 , 0.378], Fig. S9C, Table S14).

To formally test the impact of cell-type composition on sb-eQTL detection, we performed a mediation analysis, utilizing genotype interactions with estimated epithelial cell abundance as a potential mediator ((*9*), Fig. S9D). We discovered that 60 (23%) sb-eQTLs were mediated by cell-type abundances (Average Causal Mediation Effects $P \leq 0.001$, Fig. 4C) (Table S14). Mediation by other cell types cannot be excluded, particularly by immune cells: we observe that breast sb-eGenes are enriched for immunoglobulin variable chain genes (Fisher's Exact Test, OR = 12, P = 9.2e-08. In all cases, the eQTL effect size is larger in females (Table S11). Since immunoglobulin genes are mainly expressed in B cells and are among the most sex-discriminative genes in breast (Fig. S7D), we hypothesize that immunoglobulin sb-eQTLs may be driven by greater abundances of this cell type in female breasts. Collectively, these results indicate that a large proportion of sb-eQTLs in breast are driven by cell-type-specific genetic effects on gene expression that become apparent when cell types differ between sexes, although our analysis cannot distinguish whether the tested cell types or others correlated with them (Fig. S9E) are the true mediators of the signal.

Sex-aware eQTL-GWAS colocalization provides insights into the genetic basis of complex traits

To assess the utility of sb-eQTLs to dissect the molecular basis of complex trait associations, we performed colocalization (*48*) between sex-stratified *cis*-eQTLs and 87 GWASes, representing 74 distinct complex traits, for 1,089 sb-eGenes at a more relaxed FDR ($FDR \leq 0.50$) (*9*). We identified 74 colocalized gene-trait pairs (posterior probability of sharing the same causal variant, $PP4 > 0.5$; Fig. 5A). Of these, 58 were colocalized ($PP4 > 0.5$) in one sex but not in the other - 36 for females and 22 for males - corresponding to 36 unique genetic loci and 27 distinct traits (Fig. 5, A and B, Table S15). For 24/36 (67%) female- and 10/22 (45%) male-stratified *cis*-eQTL-trait pairs, evidence for colocalization was also found using the male and female combined GTEx v8 *cis*-eQTLs (Fig. S10A). For these 34 loci that colocalized in the sex-combined approach, we provide evidence that the colocalization signal is driven by regulatory effects in a single sex. The remaining 12/36 (33%) female and 12/22 (55%) male gene-trait colocalizations were not discovered with the sex-combined approach.

The strongest colocalizations between a trait and a female-stratified *cis*-eQTL were identified for *CCDC88C* and breast cancer, and for *HKDC1* and birth weight (Fig. 5C). Conversely, the strongest colocalizations between a trait and a male-stratified *cis*-eQTL were identified for *DPYSL4* and percentage of body fat, and for *CLDN7* and birth weight (Fig. 5D). *CCDC88C* is a negative regulator of the Wnt signalling pathway, a key mechanism in cancer progression (*49*) and the *CCDC88C* female *cis*-eQTL signal in breast colocalizes with risk of breast cancer (Fig. 5C, left), a trait with highly sex-differentiated incidence and presentation (*50*). For breast cancer, we identified two additional female-driven ($PP4_F > PP4_M$) colocalized sb-eGenes, *NTN4* and *CRLF3* (Table S15), previously reported as breast cancer-relevant genes (*51*, *52*).

We also discovered a preferential colocalization of blood and immune traits with female-stratified compared to male-stratified *cis*-eQTLs (odds ratio = 2.22, Fisher's exact test P = 0.047). This includes inflammatory bowel diseases, which show a higher prevalence in females with increasing age (*53*), and immune cell abundances in blood, which also exhibit sex differences (*30*, *31*). Together, these results suggest that sex-biased genetic regulation of gene expression may contribute to the etiology of diseases with marked sex differences.

Moreover, we identified colocalization signal for eQTLs and GWAS of sex-specific traits and signal possibly derived from sex-specific conditions, e.g., pregnancy in females and balding patterns in males. The *C9orf66* male-stratified *cis*-eQTL signal in breast colocalized with balding patterns in males, and the *HKDC1* female-stratified *cis*-eQTL signal in liver colocalized with birth weight, which is strongly influenced by maternal factors (Fig. 5C, right) (*54*). The sb-eQTL for this locus in liver was replicated in an independent dataset (*55*) (rs35696875-*HKDC1* P_F = 2.73e-08, P_M = 1.60e-04, z-test P = 0.004, Fig. S10B). *HKDC1* encodes a member of the hexokinase protein family and is involved in glucose metabolism. Multiple variants in perfect or high linkage disequilibrium with rs35696875 that cause reduced expression of *HKDC1* have been associated with gestational diabetes mellitus risk (*56*) and glycemic traits during pregnancy (*54*). Here, we confirm that the *HKDC1* female eQTL signal in the liver colocalizes with maternal glucose levels in plasma during pregnancy (PP4=0.92, Fig. S10C). Recently, regulatory variants spanning multiple enhancers were found to have a coordinated allelic effect on *HKDC1* expression in hepatocyte-derived cells (*57*). Estimates of hepatocyte abundance in GTEx liver samples did not differ by sex (P = 0.30) and the rs35696875-*HKDC1* sb-eQTL showed no evidence of being a hepatocyte ieQTL ($P_{G\times Hepatocytes}$ = 0.11) (*29*). Thus, unlike many sb-eQTLs in breast, the *HKDC1* sb-eQTL in liver did not seem to be driven by sex-differentiated cell-type abundances. The *HKDC1* sb-eQTL alternative allele is associated with lower *HKDC1* expression, higher maternal glucose levels and increased birth weight. These results suggest that the *HKDC1* female *cis*-eQTL influences glucose metabolism in the pregnant female, which is reflected in the birth weight of the offspring. Further investigation is needed, however, to prove causality.

Additionally, the *DPYSL4* male-stratified *cis*-eQTL signal in skeletal muscle colocalized with genetic signal associated with percentage of body fat (Fig. 5D, right). *DPYSL4* is linked to the pathophysiology of obesity and cancer: p53-inducible *DPYSL4* associates with mitochondrial supercomplexes and regulates energy metabolism in adipocytes and cancer cells. Low *DPYSL4* expression is associated with poor survival of breast cancer patients (*58*). Of note, while the colocalizing signal was detected with the male-stratified *cis*-eQTL signal, the low probability of colocalization appears to be due to the presence of an additional *cis*-eQTL in females that is absent in males. These results suggest that characterizing sex differences in the genetic associations of complex traits and molecular phenotypes can prove useful to dissect allelic heterogeneity.

For five colocalized sb-eGenes (*CLDN7*, *CCDC125*, *FAM53B*, *PLEC*, *SOWAHC*), corresponding cell-type interaction *cis*-eQTL (cell-type ieQTL) signals also colocalized with reported GWAS signals (birth weight, blood cell counts, height, platelet counts and schizophrenia, respectively) (*29*). For instance, the male-biased *cis*-eQTL rs34958987-*CLDN7* in breast (Fig. 5D, left; Fig. S10D) was identified as an epithelial cell ieQTL in breast (*29*). Both the sb-eQTL and cell-type ieQTL signals colocalize with the birth weight GWAS signal (Fig. S10E). This suggests that the origin of these sex differences in gene-trait associations may be in sex-differentiated cell-type abundances.

Finally, to assess whether sex-biased eQTL signals are reflected in sex-biased GWAS effects we obtained sex-stratified GWAS data for 36 of the 58 colocalized gene-trait pairs ((*9*), Table S15). We identified two out of 36 loci with sex differences in GWAS effect size (FDR ≤ 0.05, Bonferroni correction). These two signals correspond to *RNASET2* and *CELSR2* genes, more strongly associated to hyperthyroidism in females and to heart attack in males, respectively. However, with

the current GWAS sample sizes, we observed that, in general, sex-biased effects at the eQTL level do not readily translate into sex-biased effects at the GWAS level, in line with recent power calculations where millions of GWAS samples are estimated to be needed to address this question (*37*).

Overall, our colocalization results identified loci where sex-differentiated cell-type abundances mediate genotype-phenotype associations, and also loci where sex may play a more direct role in the underlying molecular mechanism of the association, as in the *HKDC1* locus. For future studies, accounting for context or environment (sex, in the present study) in colocalization approaches is a promising approach to discover gene-trait associations and their underlying origins.

Discussion

Here we performed an extensive characterization of sex-biased gene expression and sex differences in the genetic regulation of gene expression in *cis* across 44 human tissue sources and characterized sex differences in tissue cell-type composition for seven cell types.

We identified widespread sex-biased gene expression in all tissues, with 37% of genes exhibiting sex bias in at least one tissue, but with overall small (median FC=1.04) sex effects. These results derive from overlapping male and female distributions of interindividual expression variation, indicative of differential expression as opposed to completely dimorphic expression. These genes represent diverse molecular and biological functions, and include genes relevant to disease and clinical phenotypes. As expected, the strongest sex bias was observed for X-chromosome genes, while the vast majority of sex-biased genes were autosomal, suggesting the influence of sex on genome-wide regulatory programs. As reported in (*59*) but not well characterized to date, we discovered that a portion of these genes were non-randomly distributed across the genome, suggesting sex differences in regional regulation. Integration of these results with sex-aware analysis of epigenetic and Hi-C data may provide mechanistic insights into these patterns.

Although we identified a set of X-linked genes with sex-biased expression across many tissues, the overall tissue-sharing of sex-biased expression was strongly skewed toward tissue-specificity, with 18.2% of sex-biased genes discovered in only a single tissue. The high tissue-specificity of sex-biased gene expression and the enrichment of TFBSs in sex-biased gene promoters implicates specific TFs in mediating sex-biased expression. Functional experiments to assess sex-differentiated TF binding are needed to evaluate the role of TF function in observed patterns.

In contrast to the large impact of sex on gene expression levels, the overall extent of sex effects on genetic regulation in *cis* is much less (369 sb-eQTLs). This observation is consistent with an overall weaker role of sex on genetic regulation, but is also impacted by differences in power of the two analyses (*60*). For sb-eQTLs, the combination of small genotype-by-sex interaction effect sizes, high interindividual expression heterogeneity, and the sex imbalance in the GTEx collection affect the power of the interaction test. This implies that, to fully characterize this phenomenon, much larger cohorts are needed, particularly to assess sex effects for all *cis*-variants and genes. The relatively modest number of G×Sex interactions for a factor as impactful as sex suggests that other, more subtle genotype-interacting environmental factors are likely to be challenging to identify (as noted in (*39*)). The sb-eQTL analysis is also impacted by cell-type heterogeneity within tissues. We demonstrate that a portion of sb-eQTLs are mediated by cell-type composition,

suggesting that a portion of the sb-eQTL signal may derive from the combination of cell-type-specific eQTLs and sex differences in the tissue's cell-type composition. The remaining loci for which we had no evidence of cell-type mediation may represent true sex differences in genetic regulation of these genes, but might also derive from unknown factors confounded with sex, including cell types that were not part of our analysis. Thus, the full impact of cell-type differences across tissues remains to be determined. The identification of sb-eQTLs that are, unequivocally not derived from sex differences in cell-type abundances cannot be assessed with analysis of sb-eQTLs in bulk tissue. We anticipate that single cell sb-eQTL analysis will help disentangle sex effects on the genetics of gene expression that derive from sex differences in tissue composition versus those that derive from sex chromosome status. However, this approach also has limitations due to the removal of cells from the *in situ* tissue environment, including, for example, the presence of other cell types and diverse hormonal environments.

To understand the molecular basis of sex differences in disease and other phenotypes, it is important to note that the connection between the molecular changes observed here and complex phenotypes is likely to be complicated by many compensatory and buffering effects (*61*). Despite extensive sex differences at the transcriptome level, we highlight that the majority of biology at all phenotype levels is shared between males and females. Furthermore, the sex differences observed here are based on a snapshot of mostly older individuals. Sex differences that occur during different developmental stages, in specific environments, or in specific disease states are not well represented in our analysis. For example, sex biases are observed in many cancers (*1*). Our results provide a resource of sex effects in 'non-diseased' tissues to compare with those of disease cohorts. We note that sex is highly correlated with many features of behavior and external environments, e.g. smoking (*62*), and disentangling sex differences driven by inherent biology versus gendered environments is an important further challenge.

Beyond gene expression, sex-biased genetic regulation may also contribute to higher order phenotypes such as complex traits and diseases; colocalization analysis of sex-stratified *cis*-eQTLs and sex-combined GWAS summary statistics yielded variant-gene-trait associations that were not detected in combined-sex *cis*-eQTL colocalization analysis. In general, context-aware colocalization analyses may help to elucidate the origin of gene-trait associations, as hypothesized here for *HKDC1's* impact on birth weight through alteration of glucose metabolism in a pregnant female's liver. We show that sex-biased gene-trait associations are likely either due to allelic heterogeneity in the combined-sex cohort or genetic effects on gene expression that are (predominantly) driven by a single sex; colocalized sb-eGenes cannot be considered as proxies of loci harbouring sex differences in the genetic architecture of the linked trait. Since sex-aware colocalizations can provide insights into the sex-differentiated genetic architecture of disease, we expect future work in this area combining sex-stratified *cis*-eQTLs with summary statistics from sex-stratified GWAS to fully comprehend the impact of sex on human health and disease. Extension of analytical approaches to facilitate widespread genetic analysis of sex chromosomes is an important step to enable these new research directions.

Materials and Methods summary

Sex-differential expression was performed with *voom-limma* (*63*) and *MASH* (*64*) (Fig. S1A). Sex-differential effect sizes and gene expression levels were investigated for tissue specificity with the Tau index (*65*), clustered with *pvclust* (*66*) and compared with *dendextend* (*67*) (Fig. S3A). Sex predictivity of sex-biased genes per tissue was quantified through gradient boosted tree classifier models (*68*) (Fig. S4A). Positional gene enrichment analysis of sex-biased genes was performed with *PGE* (*18*) (Fig. S5A). Transcription factor binding site enrichment in promoter regions of sex-biased genes was performed with *Unibind (22)* and *runLOLA (69)* (Fig. S5E). Gene set enrichment analysis was performed with *fgsea (70)* (Fig. S6A) and results characterized with *Cytoscape* (*71*). Sex differences in cell type abundances and their effect on histopathological phenotypes were explored using linear regression. sb-eQTL mapping was implemented using an adaptation of *FastQTL* (*72*) (Fig. S8A); sb-eQTLs validated using haplotype-level allelic expression data generated with *phASER* and allele-specific expression modeling using *EAGLE*. Characterization of sex-specific cis-eQTL effects was performed with linear regression. Mediation of G×Sex by G×Epithelial interactions was tested with *mediation* R package. Colocalization of GWAS and eQTLs was performed with *coloc* (*48*). Further details for each analysis are provided in (*9*).

References and Notes

1. D. Zheng, J. Trynda, C. Williams, J. A. Vold, J. H. Nguyen, D. M. Harnois, S. P. Bagaria, S. A. McLaughlin, Z. Li, Sexual dimorphism in the incidence of human cancers. *BMC Cancer*. **19**, 684 (2019).

2. G. D. Anderson, Sex and racial differences in pharmacological response: where is the evidence? Pharmacogenetics, pharmacokinetics, and pharmacodynamics. *J. Womens. Health* . **14**, 19–29 (2005).

3. V. Kuan, S. Denaxas, A. Gonzalez-Izquierdo, K. Direk, O. Bhatti, S. Husain, S. Sutaria, M. Hingorani, D. Nitsch, C. A. Parisinos, R. T. Lumbers, R. Mathur, R. Sofat, J. P. Casas, I. C. K. Wong, H. Hemingway, A. D. Hingorani, A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Health*. **1**, e63–e77 (2019).

4. S. T. Ngo, F. J. Steyn, P. A. McCombe, Gender differences in autoimmune disease. *Front. Neuroendocrinol.* **35**, 347–369 (2014).

5. D. Westergaard, P. Moseley, F. K. H. Sørup, P. Baldi, S. Brunak, Population-wide analysis of differences in disease progression patterns in men and women. *Nat. Commun.* **10**, 1–14 (2019).

6. E. A. Khramtsova, L. K. Davis, B. E. Stranger, The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).

7. GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

8. F. Aguet, A. N. Barbeira, R. Bonazzola, A. Brown, S. E. Castel, B. Jo, S. Kasela, S. Kim-

Hellmuth, Y. Liang, M. Oliva, P. E. Parsana, E. Flynn, L. Fresard, E. R. Gaamzon, A. R. Hamel, Y. He, F. Hormozdiari, P. Mohammadi, M. Muñoz-Aguirre, Y. Park, A. Saha, A. V. Segrć, B. J. Strober, X. Wen, V. Wucher, S. Das, D. Garrido-Martín, N. R. Gay, R. E. Handsaker, P. J. Hoffman, S. Kashin, A. Kwong, X. Li, D. MacArthur, J. M. Rouhana, M. Stephens, E. Todres, A. Viñuela, G. Wang, Y. Zou, T. G. Consortium, C. D. Brown, N. Cox, E. Dermitzakis, B. E. Engelhardt, G. Getz, R. Guigo, S. B. Montgomery, B. E. Stranger, H. K. Im, A. Battle, K. G. Ardlie, T. Lappalainen, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, 787903 (2019).

9. "See Supplementary Materials."

10. M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, GTEx Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, R. Guigó, Human genomics. The human transcriptome across tissues and individuals. *Science*. **348**, 660–665 (2015).

11. M. Gershoni, S. Pietrokovski, The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol.* **15**, 7 (2017).

12. S. Naqvi, A. K. Godfrey, J. F. Hughes, M. L. Goodheart, R. N. Mitchell, D. C. Page, Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science*. **365** (2019), doi:10.1126/science.aaw7317.

13. C. M. Lopes-Ramos, C.-Y. Chen, M. L. Kuijjer, J. N. Paulson, A. R. Sonawane, M. Fagny, J. Platig, K. Glass, J. Quackenbush, D. L. DeMeo, Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. *Cell Rep.* **31**, 107795 (2020).

14. L. Carrel, H. F. Willard, X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*. **434**, 400–404 (2005).

15. J. L. Rinn, M. Snyder, Sexual dimorphism in mammalian gene expression. *Trends Genet.* **21**, 298–305 (2005).

16. S. García-Calzón, A. Perfilyev, V. D. de Mello, J. Pihlajamäki, C. Ling, Sex Differences in the Methylome and Transcriptome of the Human Liver and Circulating HDL-Cholesterol Levels. *J. Clin. Endocrinol. Metab.* **103**, 4395–4408 (2018).

17. S. E. Ellis, L. Collado-Torres, A. Jaffe, J. T. Leek, Improving the value of public RNA-seq expression data by phenotype prediction. *Nucleic Acids Res.* **46**, e54–e54 (2018).

18. K. De Preter, R. Barriot, F. Speleman, J. Vandesompele, Y. Moreau, Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions. *Nucleic Acids Res.* **36**, e43 (2008).

19. A. Kelkar, V. Thakur, R. Ramaswamy, D. Deobagkar, Characterisation of inactivation domains and evolutionary strata in human X chromosome through Markov segmentation. *PLoS*

One. **4**, e7885 (2009).

20. T. Tukiainen, A.-C. Villani, A. Yen, M. A. Rivas, J. L. Marshall, R. Satija, M. Aguirre, L. Gauthier, M. Fleharty, A. Kirby, B. B. Cummings, S. E. Castel, K. J. Karczewski, F. Aguet, A. Byrnes, GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, T. Lappalainen, A. Regev, K. G. Ardlie, N. Hacohen, D. G. MacArthur, Landscape of X chromosome inactivation across human tissues. *Nature*. **550**, 244–248 (2017).

21. F. Le Dily, D. Baù, A. Pohl, G. P. Vicent, F. Serra, D. Soronellas, G. Castellano, R. H. G. Wright, C. Ballare, G. Filion, M. A. Marti-Renom, M. Beato, Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.* **28**, 2151–2162 (2014).

22. M. Gheorghe, G. K. Sandve, A. Khan, J. Chèneby, B. Ballester, A. Mathelier, A map of direct TF–DNA interactions in the human genome. *Nucleic Acids Res.* **47**, e21–e21 (2019).

23. F. Spitz, E. E. M. Furlong, Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

24. N. Leuenberger, S. Pradervand, W. Wahli, Sumoylated PPARalpha mediates sex-specific gene repression and protects the liver from estrogen-induced toxicity in mice. *J. Clin. Invest.* **119**, 3138–3148 (2009).

25. N. Brockdorff, Polycomb complexes in X chromosome inactivation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372** (2017), doi:10.1098/rstb.2017.0021.

26. B. M. Nugent, C. M. O'Donnell, C. N. Epperson, T. L. Bale, Placental H3K27me3 establishes female resilience to prenatal insults. *Nat. Commun.* **9**, 2555 (2018).

27. D. Lau-Corona, W. K. Bae, L. Hennighausen, D. J. Waxman, Sex-biased genetic programs in liver metabolism and liver fibrosis are controlled by EZH1 and EZH2. *bioRxiv*, 577056 (2019).

28. D. F. Deegan, N. Engel, Sexual Dimorphism in the Age of Genomics: How, When, Where. *Front Cell Dev Biol*. **7**, 186 (2019).

29. S. Kim-Hellmuth, F. Aguet, M. Oliva, M. Muñoz-Aguirre, V. Wucher, S. Kasela, S. E. Castel, A. R. Hamel, A. Viñuela, A. L. Roberts, S. Mangul, X. Wen, G. Wang, A. N. Barbeira, D. Garrido-Martín, B. Nadel, Y. Zou, R. Bonazzola, J. Quan, A. Brown, A. Martinez-Perez, J. M. Soria, G. Consortium, G. Getz, E. T. Dermitzakis, K. S. Small, M. Stephens, H. S. Xi, H. K. Im, R. Guigó, A. V. Segrè, B. E. Stranger, K. G. Ardlie, T. Lappalainen, Cell type specific genetic

regulation of gene expression across human tissues. *bioRxiv*, 806117 (2019).

30. Y. Chen, Y. Zhang, G. Zhao, C. Chen, P. Yang, S. Ye, X. Tan, Difference in Leukocyte Composition between Women before and after Menopausal Age, and Distinct Sexual Dimorphism. *PLoS One*. **11**, e0162953–e0162953 (2016).

31. E. Bongen, H. Lucian, A. Khatri, G. K. Fragiadakis, Z. B. Bjornson, G. P. Nolan, P. J. Utz, P. Khatri, Sex Differences in the Blood Transcriptome Identify Robust Changes in Immune Cell Proportions with Aging and Influenza Infection. *Cell Rep.* **29**, 1961–1973.e4 (2019).

32. W. J. Astle, H. Elding, T. Jiang, D. Allen, D. Ruklisa, A. L. Mann, D. Mead, H. Bouman, F. Riveros-Mckay, M. A. Kostadima, J. J. Lambourne, S. Sivapalaratnam, K. Downes, K. Kundu, L. Bomba, K. Berentsen, J. R. Bradley, L. C. Daugherty, O. Delaneau, K. Freson, S. F. Garner, L. Grassi, J. Guerrero, M. Haimel, E. M. Janssen-Megens, A. Kaan, M. Kamat, B. Kim, A. Mandoli, J. Marchini, J. H. A. Martens, S. Meacham, K. Megy, J. O'Connell, R. Petersen, N. Sharifi, S. M. Sheard, J. R. Staley, S. Tuna, M. van der Ent, K. Walter, S.-Y. Wang, E. Wheeler, S. P. Wilder, V. Iotchkova, C. Moore, J. Sambrook, H. G. Stunnenberg, E. Di Angelantonio, S. Kaptoge, T. W. Kuijpers, E. Carrillo-de-Santa-Pau, D. Juan, D. Rico, A. Valencia, L. Chen, B. Ge, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yang, R. Guigo, S. Beck, D. S. Paul, T. Pastinen, D. Bujold, G. Bourque, M. Frontini, J. Danesh, D. J. Roberts, W. H. Ouwehand, A. S. Butterworth, N. Soranzo, The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. **167**, 1415–1429.e19 (2016).

33. K. Rawlik, O. Canela-Xandri, A. Tenesa, Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome Biol.* **17**, 166 (2016).

34. D. Shungin, T. W. Winkler, D. C. Croteau-Chonka, T. Ferreira, A. E. Locke, R. Mägi, R. J. Strawbridge, T. H. Pers, K. Fischer, A. E. Justice, T. Workalemahu, J. M. W. Wu, M. L. Buchkovich, N. L. Heard-Costa, T. S. Roman, A. W. Drong, C. Song, S. Gustafsson, F. R. Day, T. Esko, T. Fall, Z. Kutalik, J. Luan, J. C. Randall, A. Scherag, S. Vedantam, A. R. Wood, J. Chen, R. Fehrmann, J. Karjalainen, B. Kahali, C.-T. Liu, E. M. Schmidt, D. Absher, N. Amin, D. Anderson, M. Beekman, J. L. Bragg-Gresham, S. Buyske, A. Demirkan, G. B. Ehret, M. F. Feitosa, A. Goel, A. U. Jackson, T. Johnson, M. E. Kleber, K. Kristiansson, M. Mangino, I. Mateo Leach, C. Medina-Gomez, C. D. Palmer, D. Pasko, S. Pechlivanis, M. J. Peters, I. Prokopenko, A. Stančáková, Y. Ju Sung, T. Tanaka, A. Teumer, J. V. Van Vliet-Ostaptchouk, L. Yengo, W. Zhang, E. Albrecht, J. Ärnlöv, G. M. Arscott, S. Bandinelli, A. Barrett, C. Bellis, A. J. Bennett, C. Berne, M. Blüher, S. Böhringer, F. Bonnet, Y. Böttcher, M. Bruinenberg, D. B. Carba, I. H. Caspersen, R. Clarke, E. Warwick Daw, J. Deelen, E. Deelman, G. Delgado, A. S. F. Doney, N. Eklund, M. R. Erdos, K. Estrada, E. Eury, N. Friedrich, M. E. Garcia, V. Giedraitis, B. Gigante, A. S. Go, A. Golay, H. Grallert, T. B. Grammer, J. Gräßler, J. Grewal, C. J. Groves, T. Haller, G. Hallmans, C. A. Hartman, M. Hassinen, C. Hayward, K. Heikkilä, K.-H. Herzig, Q. Helmer, H. L. Hillege, O. Holmen, S. C. Hunt, A. Isaacs, T. Ittermann, A. L. James, I. Johansson, T. Juliusdottir, I.-P. Kalafati, L. Kinnunen, W. Koenig, I. K. Kooner, W. Kratzer, C. Lamina, K. Leander, N. R. Lee, P. Lichtner, L. Lind, J. Lindström, S. Lobbens, M. Lorentzon, F. Mach, P. K. E. Magnusson, A. Mahajan, W. L. McArdle, C. Menni, S. Merger, E. Mihailov, L. Milani, R. Mills, A. Moayyeri, K. L. Monda, S. P. Mooijaart, T. W. Mühleisen, A. Mulas, G. Müller, M. Müller-Nurasyid, R.

Nagaraja, M. A. Nalls, N. Narisu, N. Glorioso, I. M. Nolte, M. Olden, N. W. Rayner, F. Renstrom, J. S. Ried, N. R. Robertson, L. M. Rose, S. Sanna, H. Scharnagl, S. Scholtens, B. Sennblad, T. Seufferlein, C. M. Sitlani, A. Vernon Smith, K. Stirrups, H. M. Stringham, J. Sundström, M. A. Swertz, A. J. Swift, A.-C. Syvänen, B. O. Tayo, B. Thorand, G. Thorleifsson, A. Tomaschitz, C. Troffa, F. V. A. van Oort, N. Verweij, J. M. Vonk, L. L. Waite, R. Wennauer, T. Wilsgaard, M. K. Wojczynski, A. Wong, Q. Zhang, J. Hua Zhao, E. P. Brennan, M. Choi, P. Eriksson, L. Folkersen, A. Franco-Cereceda, A. G. Gharavi, Å. K. Hedman, M.-F. Hivert, J. Huang, S. Kanoni, F. Karpe, S. Keildson, K. Kiryluk, L. Liang, R. P. Lifton, B. Ma, A. J. McKnight, R. McPherson, A. Metspalu, J. L. Min, M. F. Moffatt, G. W. Montgomery, J. M. Murabito, G. Nicholson, D. R. Nyholt, C. Olsson, J. R. B. Perry, E. Reinmaa, R. M. Salem, N. Sandholm, E. E. Schadt, R. A. Scott, L. Stolk, E. E. Vallejo, H.-J. Westra, K. T. Zondervan, The ADIPOGen Consortium, The CARDIOGRAMplusC4D Consortium, The CKDGen Consortium, The GEFOS Consortium, The GENIE Consortium, The Glgc, The Icbp, The International Endogene Consortium, The LifeLines Cohort Study, The MAGIC Investigators, The MuTHER Consortium, The PAGE Consortium, The ReproGen Consortium, P. Amouyel, D. Arveiler, S. J. L. Bakker, J. Beilby, R. N. Bergman, J. Blangero, M. J. Brown, M. Burnier, H. Campbell, A. Chakravarti, P. S. Chines, S. Claudi-Boehm, F. S. Collins, D. C. Crawford, J. Danesh, U. de Faire, E. J. C. de Geus, M. Dörr, R. Erbel, J. G. Eriksson, M. Farrall, E. Ferrannini, J. Ferrières, N. G. Forouhi, T. Forrester, O. H. Franco, R. T. Gansevoort, C. Gieger, V. Gudnason, C. A. Haiman, T. B. Harris, A. T. Hattersley, M. Heliövaara, A. A. Hicks, A. D. Hingorani, W. Hoffmann, A. Hofman, G. Homuth, S. E. Humphries, E. Hyppönen, T. Illig, M.-R. Jarvelin, B. Johansen, P. Jousilahti, A. M. Jula, J. Kaprio, F. Kee, S. M. Keinanen-Kiukaanniemi, J. S. Kooner, C. Kooperberg, P. Kovacs, A. T. Kraja, M. Kumari, K. Kuulasmaa, J. Kuusisto, T. A. Lakka, C. Langenberg, L. Le Marchand, T. Lehtimäki, V. Lyssenko, S. Männistö, A. Marette, T. C. Matise, C. A. McKenzie, B. McKnight, A. W. Musk, S. Möhlenkamp, A. D. Morris, M. Nelis, C. Ohlsson, A. J. Oldehinkel, K. K. Ong, L. J. Palmer, B. W. Penninx, A. Peters, P. P. Pramstaller, O. T. Raitakari, T. Rankinen, D. C. Rao, T. K. Rice, P. M. Ridker, M. D. Ritchie, I. Rudan, V. Salomaa, N. J. Samani, J. Saramies, M. A. Sarzynski, P. E. H. Schwarz, A. R. Shuldiner, J. A. Staessen, V. Steinthorsdottir, R. P. Stolk, K. Strauch, A. Tönjes, A. Tremblay, E. Tremoli, M.-C. Vohl, U. Völker, P. Vollenweider, J. F. Wilson, J. C. Witteman, L. S. Adair, M. Bochud, B. O. Boehm, S. R. Bornstein, C. Bouchard, S. Cauchi, M. J. Caulfield, J. C. Chambers, D. I. Chasman, R. S. Cooper, G. Dedoussis, L. Ferrucci, P. Froguel, H.-J. Grabe, A. Hamsten, J. Hui, K. Hveem, K.-H. Jöckel, M. Kivimaki, D. Kuh, M. Laakso, Y. Liu, W. März, P. B. Munroe, I. Njølstad, B. A. Oostra, C. N. A. Palmer, N. L. Pedersen, M. Perola, L. Pérusse, U. Peters, C. Power, T. Quertermous, R. Rauramaa, F. Rivadeneira, T. E. Saaristo, D. Saleheen, J. Sinisalo, P. Eline Slagboom, H. Snieder, T. D. Spector, U. Thorsteinsdottir, M. Stumvoll, J. Tuomilehto, A. G. Uitterlinden, M. Uusitupa, P. van der Harst, G. Veronesi, M. Walker, N. J. Wareham, H. Watkins, H.-E. Wichmann, G. R. Abecasis, T. L. Assimes, S. I. Berndt, M. Boehnke, I. B. Borecki, P. Deloukas, L. Franke, T. M. Frayling, L. C. Groop, D. J. Hunter, R. C. Kaplan, J. R. O'Connell, L. Qi, D. Schlessinger, D. P. Strachan, K. Stefansson, C. M. van Duijn, C. J. Willer, P. M. Visscher, J. Yang, J. N. Hirschhorn, M. Carola Zillikens, M. I. McCarthy, E. K. Speliotes, K. E. North, C. S. Fox, I. Barroso, P. W. Franks, E. Ingelsson, I. M. Heid, R. J. F. Loos, L. A. Cupples, A. P. Morris, C. M. Lindgren, K. L. Mohlke, New genetic loci link adipose and insulin biology to body fat distribution. *Nature*. **518**, 187–196 (2015).

35. S. L. Pulit, C. Stoneman, A. P. Morris, A. R. Wood, C. A. Glastonbury, J. Tyrrell, L. Yengo, T. Ferreira, E. Marouli, Y. Ji, J. Yang, S. Jones, R. Beaumont, D. C. Croteau-Chonka, T.

W. Winkler, G. Consortium, A. T. Hattersley, R. J. F. Loos, J. N. Hirschhorn, P. M. Visscher, T. M. Frayling, H. Yaghootkar, C. M. Lindgren, Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

36. J. Martin, E. A. Khramtsova, S. B. Goleva, G. A. M. Blokland, M. Traglia, R. K. Walters, C. Hübel, J. R. I. Coleman, G. Breen, A. D. Børglum, D. Demontis, J. Grove, T. Werge, J. Bralten, C. M. Bulik, P. H. Lee, C. A. Mathews, R. E. Peterson, S. J. Winham, N. Wray, H. J. Edenberg, W. Guo, Y. Yao, B. M. Neale, S. V. Faraone, T. L. Petryshen, L. A. Weiss, L. E. Duncan, Sex Differences Cross-Disorder Analysis Group of the Psychiatric Genomics Consortium, J. M. Goldstein, J. W. Smoller, B. E. Stranger, L. K. Davis, Examining sex-differentiated genetic effects across neuropsychiatric and behavioral traits. *bioRxiv* (2020), p. 2020.05.04.076042.

37. E. Porcu, A. Claringbould, K. Lepik, BIOS Consortium, T. G. Richardson, F. A. Santoni, L. Franke, A. Reymond, Z. Kutalik, The role of gene expression on human sexual dimorphism: too early to call. *bioRxiv* (2020), p. 2020.04.15.042986.

38. S. E. Castel, F. Aguet, P. Mohammadi, G. Consortium, K. G. Ardlie, T. Lappalainen, A vast resource of allelic expression data spanning human tissues. *bioRxiv*, 792911 (2019).

39. D. A. Knowles, J. R. Davis, H. Edgington, A. Raj, M.-J. Favé, X. Zhu, J. B. Potash, M. M. Weissman, J. Shi, D. F. Levinson, P. Awadalla, S. Mostafavi, S. B. Montgomery, A. Battle, Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods*. **14**, 699–702 (2017).

40. K. R. Kukurba, P. Parsana, B. Balliu, K. S. Smith, Z. Zappala, D. A. Knowles, M.-J. Favé, J. R. Davis, X. Li, X. Zhu, J. B. Potash, M. M. Weissman, J. Shi, A. Kundaje, D. F. Levinson, P. Awadalla, S. Mostafavi, A. Battle, S. B. Montgomery, Impact of the X Chromosome and sex on regulatory variation. *Genome Res.*, gr.197897.115 (2016).

41. The GAIT2 project, (available at http://ugcd.github.io/pages/projects/gait2/).

42. J. J. Shen, Y.-F. Wang, W. Yang, Sex-Interacting mRNA- and miRNA-eQTLs and Their Implications in Gene Expression Regulation and Disease. *Front. Genet.* **10** (2019), doi:10.3389/fgene.2019.00313.

43. C. Yao, R. Joehanes, A. D. Johnson, T. Huan, T. Esko, S. Ying, J. E. Freedman, J. Murabito, K. L. Lunetta, A. Metspalu, P. J. Munson, D. Levy, Sex- and age-interacting eQTLs in human complex diseases. *Hum. Mol. Genet.* **23**, 1947–1956 (2014).

44. A. C. Leon, M. Heo, Sample Sizes Required to Detect Interactions between Two Binary Fixed-Effects in a Mixed-Effects Linear Regression Model. *Comput. Stat. Data Anal.* **53**, 603–608 (2009).

45. K. Iida, V. Nussenzweig, Complement receptor is an inhibitor of the complement cascade. *J. Exp. Med.* **153**, 1138–1150 (1981).

46. K. Dhanasekaran, A. Bose, V. J. Rao, R. Boopathi, S. R. Shankar, V. K. Rao, A. Swaminathan, M. Vasudevan, R. Taneja, T. K. Kundu, Unraveling the role of aurora A beyond centrosomes and spindle assembly: implications in muscle differentiation. *FASEB J.* **33**, 219–230 (2019).

47. A. Tang, K. Gao, L. Chu, R. Zhang, J. Yang, J. Zheng, Aurora kinases: novel therapy targets in cancers. *Oncotarget*. **8**, 23937–23954 (2017).

48. C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, V. Plagnol, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

49. T. Zhan, N. Rindtorff, M. Boutros, Wnt signaling in cancer. *Oncogene*. **36**, 1461–1473 (2017).

50. D. Ly, D. Forman, J. Ferlay, L. A. Brinton, M. B. Cook, An international comparison of male and female breast cancer incidence rates. *Int. J. Cancer*. **132**, 1918–1926 (2013).

51. X. Xu, Q. Yan, Y. Wang, X. Dong, NTN4 is associated with breast cancer metastasis via regulation of EMT-related biomarkers. *Oncol. Rep.* **37**, 449–457 (2017).

52. M. M. Marjaneh, H. Sivakumaran, K. M. Hillman, S. Kaufmann, N. Hussein, L. G. Lima, S. Ham, S. Kar, J. Beesley, L. Fachal, D. F. Easton, A. M. Dunning, A. Möller, G. Chenevix-Trench, S. L. Edwards, J. D. French, High-throughput allelic expression imbalance analyses identify candidate breast cancer risk genes. *bioRxiv*, 521013 (2019).

53. J. D. Betteridge, S. P. Armbruster, C. Maydonovitch, G. R. Veerappan, Inflammatory bowel disease prevalence by age, gender, race, and geographic location in the U.S. military health care population. *Inflamm. Bowel Dis.* **19**, 1421–1427 (2013).

54. M. G. Hayes, M. Urbanek, M.-F. Hivert, L. L. Armstrong, J. Morrison, C. Guo, L. P. Lowe, D. A. Scheftner, A. Pluzhnikov, D. M. Levine, C. P. McHugh, C. M. Ackerman, L. Bouchard, D. Brisson, B. T. Layden, D. Mirel, K. F. Doheny, M. V. Leya, R. N. Lown-Hecht, A. R. Dyer, B. E. Metzger, T. E. Reddy, N. J. Cox, W. L. Lowe, HAPO Study Cooperative Research Group, Identification of HKDC1 and BACE2 as genes influencing glycemic traits during pregnancy through genome-wide association studies. *Diabetes*. **62**, 3282–3291 (2013).

55. T. Strunz, F. Grassmann, J. Gayán, S. Nahkuri, D. Souza-Costa, C. Maugeais, S. Fauser, E. Nogoceke, B. H. F. Weber, A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci. Rep.* **8** (2018), doi:10.1038/s41598-018-24219-z.

56. W. L. Lowe, D. M. Scholtens, V. Sandler, M. G. Hayes, Genetics of Gestational Diabetes Mellitus and Maternal Metabolism. *Curr. Diab. Rep.* **16**, 15 (2016).

57. C. Guo, A. E. Ludvik, M. E. Arlotto, M. G. Hayes, L. L. Armstrong, D. M. Scholtens, C. D. Brown, C. B. Newgard, T. C. Becker, B. T. Layden, W. L. Lowe, T. E. Reddy, Coordinated

regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase HKDC1. *Nat. Commun.* **6**, 6069 (2015).

58. H. Nagano, N. Hashimoto, A. Nakayama, S. Suzuki, Y. Miyabayashi, A. Yamato, S. Higuchi, M. Fujimoto, I. Sakuma, M. Beppu, M. Yokoyama, Y. Suzuki, S. Sugano, K. Ikeda, I. Tatsuno, I. Manabe, K. Yokote, S. Inoue, T. Tanaka, p53-inducible DPYSL4 associates with mitochondrial supercomplexes and regulates energy metabolism in adipocytes and cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 8370–8375 (2018).

59. B. J. Matthews, D. J. Waxman, CTCF and Cohesin link sex-biased distal regulatory elements to sex-biased gene expression in mouse liver. *bioRxiv* (2019) (available at https://www.biorxiv.org/content/10.1101/577577v1.abstract).

60. G. Shieh, Effect size, statistical power, and sample size for assessing interactions between categorical and continuous variables. *Br. J. Math. Stat. Psychol.* **72**, 136–154 (2019).

61. Y. Liu, A. Beyer, R. Aebersold, On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*. **165**, 535–550 (2016).

62. A. M. Allen, T. S. Scheuermann, N. Nollen, D. Hatsukami, J. S. Ahluwalia, Gender Differences in Smoking Behavior and Dependence Motives Among Daily and Nondaily Smokers. *Nicotine Tob. Res.* **18**, 1408–1413 (2016).

63. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

64. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).

65. N. Kryuchkova-Mostacci, M. Robinson-Rechavi, A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).

66. R. Suzuki, H. Shimodaira, *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling* (2015).

67. T. Galili, dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*. **31**, 3718–3720 (2015).

68. T. Chen, C. Guestrin, in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2016), *KDD '16*, pp. 785–794.

69. N. C. Sheffield, C. Bock, LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* (2016) (available at http://code.databio.org/LOLA).

70. A. A. Sergushichev, An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, 060012 (2016).

71. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

72. H. Ongen, A. Buil, A. A. Brown, E. T. Dermitzakis, O. Delaneau, Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*. **32**, 1479–1485 (2016).

73. D. Aran, Z. Hu, A. J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).

74. J. Chen, E. Behnam, J. Huang, M. F. Moffatt, D. J. Schaid, L. Liang, X. Lin, Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. *BMC Genomics*. **18**, 413 (2017).

75. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–9445 (2003).

76. M. Stephens, False discovery rates: a new deal. *Biostatistics*. **18**, 275–294 (2017).

77. M. Allen, M. M. Carrasquillo, C. Funk, B. D. Heavner, F. Zou, C. S. Younkin, J. D. Burgess, H.-S. Chai, J. Crook, J. A. Eddy, H. Li, B. Logsdon, M. A. Peters, K. K. Dang, X. Wang, D. Serie, C. Wang, T. Nguyen, S. Lincoln, K. Malphrus, G. Bisceglio, M. Li, T. E. Golde, L. M. Mangravite, Y. Asmann, N. D. Price, R. C. Petersen, N. R. Graff-Radford, D. W. Dickson, S. G. Younkin, N. Ertekin-Taner, Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data*. **3**, 160089 (2016).

78. G. Stone, A. Choi, O. Meritxell, J. Gorham, M. Heydarpour, C. E. Seidman, J. G. Seidman, S. F. Aranki, S. C. Body, V. J. Carey, B. A. Raby, B. E. Stranger, J. D. Muehlschlegel, Sex differences in gene expression in response to ischemia in the human left ventricular myocardium. *Hum. Mol. Genet.* **28**, 1682–1693 (2019).

79. T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, E. T. Dermitzakis, Geuvadis Consortium, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. **501**, 506–511 (2013).

80. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn:

Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2012).

81. S. M. Lundberg, G. G. Erion, S.-I. Lee, Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802. 03888 [cs, stat]* (2018) (available at http://arxiv.org/abs/1802.03888).

82. A. M. Cotton, E. M. Price, M. J. Jones, B. P. Balaton, M. S. Kobor, C. J. Brown, Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* **24**, 1528–1539 (2015).

83. K. Wainer Katsir, M. Linial, Human genes escaping X-inactivation revealed by single cell expression data. *BMC Genomics*. **20**, 201 (2019).

84. Y. Zhang, A. Castillo-Morales, M. Jiang, Y. Zhu, L. Hu, A. O. Urrutia, X. Kong, L. D. Hurst, Genes That Escape X-Inactivation in Humans Have High Intraspecific Variability in Expression, Are Associated with Mental Impairment but Are Not Slow Evolving. *Mol. Biol. Evol.* **30**, 2588–2601 (2013).

85. M. D. Schultz, Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang, J. R. Ecker, Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*. **523**, 212–216 (2015).

86. C. Park, L. Carrel, K. D. Makova, Strong Purifying Selection at Genes Escaping X Chromosome Inactivation. *Mol. Biol. Evol.* **27**, 2446–2450 (2010).

87. B. J. Schmiedel, D. Singh, A. Madrigal, A. G. Valdovino-Gonzalez, B. M. White, J. Zapardiel-Gonzalo, B. Ha, G. Altay, J. A. Greenbaum, G. McVicker, G. Seumois, A. Rao, M. Kronenberg, B. Peters, P. Vijayanand, Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*. **175**, 1701–1715.e16 (2018).

88. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

89. B. T. Lahn, Four Evolutionary Strata on the Human X Chromosome. *Science*. **286**, 964–967 (1999).

90. P. Pihlajamaa, B. Sahu, O. A. Jänne, Determinants of Receptor- and Tissue-Specific Actions in Androgen Signaling. *Endocr. Rev.* **36**, 357–384 (2015).

91. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

92. H. Araki, C. Knapp, P. Tsai, C. Print, GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*. **2**, 76–82 (2012).

93. K. Slowikowski, tftargets (2017), (available at https://github.com/slowkow/tftargets).

94. H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, E. Kim, S. Lee, B. Kang, D. Jeong, Y. Kim, H.-N. Jeon, H. Jung, S. Nam, M. Chung, J.-H. Kim, I. Lee, TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).

95. A. Breschi, M. Muñoz-Aguirre, V. Wucher, C. A. Davis, D. Garrido-Martín, S. Djebali, J. Gillis, D. D. Pervouchine, A. Vlasova, A. Dobin, C. Zaleski, J. Drenkow, C. Danyko, A. Scavelli, F. Reverter, M. P. Snyder, T. R. Gingeras, R. Guigó, A limited set of transcriptional programs define major cell types. *bioRxiv* (2020), p. 857169.

96. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).

97. P. Mohammadi, S. E. Castel, A. A. Brown, T. Lappalainen, Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).

98. S. E. Castel, P. Mohammadi, W. K. Chung, Y. Shen, T. Lappalainen, Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.* **7**, 12817 (2016).

99. D. A. Knowles, C. K. Burrows, J. D. Blischak, K. M. Patterson, D. J. Serie, N. Norton, C. Ober, J. K. Pritchard, Y. Gilad, Determining the genetic basis of anthracycline-cardiotoxicity by molecular response QTL mapping in induced cardiomyocytes. *Elife*. **7** (2018), doi:10.7554/eLife.33480.

100. A. Saha, A. Battle, False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res.* **7**, 1860 (2018).

101. A. N. Barbeira, R. Bonazzola, E. R. Gamazon, Y. Liang, Y. Park, S. Kim-Hellmuth, G. Wang, Z. Jiang, D. Zhou, F. Hormozdiari, B. Liu, A. Rao, A. R. Hamel, M. D. Pividori, F. Aguet, GTEx GWAS Working Group, L. Bastarache, D. M. Jordan, M. Verbanck, R. Do, GTEx Consortium, M. Stephens, K. Ardlie, M. McCarthy, S. B. Montgomery, A. V. Segrè, C. D. Brown, T. Lappalainen, X. Wen, H. K. Im, Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits. *bioRxiv* (2019), p. 814350.

102. X. Wen, R. Pique-Regi, F. Luca, Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).

Acknowledgments:

We thank the donors and their families for their generous gifts of biospecimens to the GTEx research project; the Genomics Platform at the Broad Institute for data generation; J. Struewing for his support and leadership of the GTEx project; B. HF Weber and T. Strunz for assistance with replicating the sb-eQTL for *HKDC1* in liver; D. Nicolae and L. Chen for advice on mediation analysis, J. Witkos for comments on an earlier version of the manuscript, and M. Gloudemans for making the sex-stratified eQTL data available on LocusCompare (http://locuscompare.com/), D. Muehlschlegel for assistance with replicating sex-biased genes, E. Flynn for assistance in the interpretation of sex-biased gene expression patterns, and G. Hayes for providing GWAS summary statistics for maternal glycemic traits. The article cover summary Figure and Figure 1 were partially generated using https://www.biorender.com, and M. Khan and C. Stolte contributed to the design. This work was completed in part with computational resources provided by the Center for Research Informatics at The University of Chicago, and the Centre for Genomic Regulation. The Center for Research Informatics is funded by the Biological Sciences Division at the University of Chicago with additional funding provided by the Institute for Translational Medicine, CTSA grant number UL1 TR000430 from the National Institutes of Health.

Funding: This work was supported by the Common Fund of the Office of the Director, U.S. National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, NIA, NIAID, and NINDS through NIH contracts HHSN261200800001E (Leidos Prime contract with NCI: A.M.S., D.E.T., N.V.R., J.A.M., L.S., M.E.B., L.Q., T.K., D.B., K.R., A.U.), 10XS170 (NDRI: W.F.L., J.A.T., G.K., A.M., S.S., R.H., G.Wa., M.J., M.Wa., L.E.B., C.J., J.W., B.R., M.Hu., K.M., L.A.S., H.M.G., M.Mo., L.K.B.), 10XS171 (Roswell Park Cancer Institute: B.A.F., M.T.M., E.K., B.M.G., K.D.R., J.B.), 10X172 (Science Care Inc.), 12ST1039 (IDOX), 10ST1035 (Van Andel Institute: S.D.J., D.C.R., D.R.V.), HHSN268201000029C (Broad Institute: F.A., G.G., K.G.A., A.V.S., X.Li., E.T., S.G., A.G., S.A., K.H.H., D.T.N., K.H., S.R.M., J.L.N.), 5U41HG009494 (F.A., G.G., K.G.A.), and through NIH grants R01 DA006227-17 (Univ. of Miami Brain Bank: D.C.M., D.A.D.), Supplement to University of Miami grant DA006227 (D.C.M., D.A.D.), R01 MH090941 (Univ. of Geneva), R01 MH090951 and R01 MH090937 (Univ. of Chicago), R01 MH090936 (Univ. of North Carolina–Chapel Hill), R01MH101814 (M.M-A., V.W., S.B.M., R.G., E.T.D., D.G-M., A.V.), U01HG007593 (S.B.M.), R01MH101822 (C.D.B.), U01HG007598 (M.O., B.E.S.), U01MH104393 (A.P.F.), extension H002371 to 5U41HG002371 (W.J.K) as well as other funding sources: R01MH106842 (T.L., P.M., E.F., P.J.H.), R01HL142028 (T.L., Si.Ka., P.J.H.), R01GM122924 (T.L., S.E.C.), R01MH107666 (H.K.I.), P30DK020595 (H.K.I.), UM1HG008901 (T.L.), R01GM124486 (T.L.), R01HG010067 (Y.Pa.), R01HG002585 (G.Wa., M.St.), Gordon and Betty Moore Foundation GBMF 4559 (G.Wa., M.St.), 1K99HG009916-01 (S.E.C.), R01HG006855 (Se.Ka., R.E.H.), BIO2015-70777-P, Ministerio de Economía y Competitividad and FEDER funds (M.M-A., V.W., R.G., D.G-M.), la Caixa Foundation ID 100010434 under agreement LCF/BQ/SO15/52260001 (D.G-M.), NIH CTSA grant UL1TR002550-01 (P.M.), Marie-Skłodowska Curie fellowship H2020 Grant 706636 (S.K-H.), R35HG010718 (E.R.G.), FPU15/03635, Ministerio de Educación, Cultura y Deporte (M.M-A.), R01MH109905, 1R01HG010480 (A.Ba.), Searle Scholar Program (A.Ba.), R01HG008150 (S.B.M.), 5T32HG000044-22, NHGRI Institutional Training Grant in Genome Science (N.R.G.), EU IMI program (UE7-DIRECT-115317-1) (E.T.D., A.V.), FNS funded project RNA1 (31003A_149984) (E.T.D., A.V.), DK110919 (F.H.), F32HG009987 (F.H.), Massachusetts Lions

Eye Research Fund Grant (A.R.H.), 2R01GM108711 (L.C.), R01MH101820 (B.E.S.), Supplement to R01MH101820 (E.A.K., P.E., B.E.S), Consolidate Research Group. Generalitat de Catalunya SGR 1736 and CERCA program (A.M.-P., J.M.S.); Rhodes Trust and Natural Sciences and Engineering Research Council of Canada (A.J.P.). All CRG authors acknowledge the support of the Spanish Ministry of Science, Innovation and Universities to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme / Generalitat de Catalunya.

Author contributions: B.E.S. conceived the study; M.O. and B.E.S. led the writing and editing of the manuscript and supplement; M.O. and B.E.S. coordinated analyses of all contributing authors; M.O., M.M.-A. and V.W. performed differential gene expression analysis; B.B., D.J.C., M.M.-A., M.O., V.W. characterized effect sizes of sex-biased genes; M.M.-A and M.O., performed sex-biased genes replication in independent datasets, M.M.-A., M.O., and V.W performed analysis of transcription factor binding sites; M.M.-A., M.O., and V.W performed tissue clustering based on gene expression levels and sex bias; M.M.-A. built the expression-based sex classifier; Y.Z. performed *MASH* analyses; M.S. and S.K.-H. provided advice on *MASH* analysis; M.O. generated sb-eQTL pipelines and performed sb-eQTL mapping; F.A., B.B., A.J.B., B.E.E., E.E., P.E., E.R.G., S.K.-H., S.K., E.A.K., S.B.M., P.P., A.D.S., and B.E.S. contributed to sb-eQTL analysis approach; D.J.C., S.K.-H., E.A.K., M.O., and V.W. characterized sb-eGenes; F.A., B.B. and A.V. performed sb-eQTL replication analysis in external datasets; S.K.-H., A.M.-P. and J.-M.S. contributed to sb-eQTL replication analysis; S.K. performed ASE aFC validation of sb-eQTLs; S.E.C., S.K.-H. and P.M. contributed to ASE aFC validation of sb-eQTLs; P.P. performed EAGLE ASE validation of sb-eQTLs; A.J.B. and S.K.-H. provided advice on EAGLE ASE validation; S.K.-H. performed coloc analysis; M.O. performed mediation analysis; S.K.-H. and B.L.P. provided advice on mediation analysis; F.A., D.G., S.K.-H., D.J.C., M.O., M.M.-A, and V.W. generated figures; F.A., K.G.A., and A.V.S. generated and oversaw GTEx v8 data generation, LDACC & pipelines; A.N.B., R.B., and H.K.I. generated GWAS data; F.A., S.K.-H., and M.O. generated cell-type abundances and ieQTL data; S.K.-H., M.M.-A., M.O. characterized sex differences in cell-type abundances; M.M.-A. and V.W. characterized phenotype relationships with cell-type abundances; A.B., A.D.H.G., A.R.H., E.A.K., A.J.P., B.E.E, D.G., E.R.G., S.K.-H., A.M.-P., F.R., and A.D.S. performed analysis or provided feedback that significantly shaped this work but was not included in this final version; M.M.-A. and V.W. managed data and code in the GitHub repository; A.J.B., B.E.E., E.T.D., R.G., H.K.I., T.L., S.B.M., B.L.P., M.S., A.V.S., and B.E.S. supervised the work of trainees in his/her lab; D.J.C., S.K., S.K.-H., M.M.-A., M.O., P.P., V.W., Y.Z., and B.E.S. wrote manuscript text; B.B., A.J.B., B.E.E., A.D.H.G., R.G., S.K.-H., H.K.I., E.A.K., T.L., M.M.-A., M.O., S.B.M., L.C., S.K., P.P., B.L.P., A.D.S., B.E.S., A.V., and V.W., edited the manuscript. All authors read and approved the final manuscript.

Competing interests: F.A. is an inventor on a patent application related to TensorQTL; S.E.C. is a co-founder, chief technology officer and stock owner at Variant Bio; D.G.M. is co-founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme; E.A.K. is an employee of Janssen Pharmaceuticals; H.I. has received speaker honoraria from GSK and AbbVie; E.T.D. is chairman and member of the board of Hybridstat LTD; T.L. is a scientific advisory board member of Variant Bio with equity, and Goldfinch Bio. Other GTEx members: E.R.G. is on the Editorial Board of Circulation Research, and does consulting for the City of Hope / Beckman Research Institute;

E.T.D. is chairman and member of the board of Hybridstat LTD.; B.E.E. is on the scientific advisory boards of Celsius Therapeutics and Freenome; G.G. receives research funds from IBM and Pharmacyclics, and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect, MSMutSig, POLYSOLVER and TensorQTL. G.G. is a founder, consultant and holds privately held equity in Scorpion Therapeutics; S.B.M. is on the scientific advisory board of MyOme; P.F. is member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomes, Ltd. P.G.F. is a partner of Bioinf2Bio.

Data and materials availability: All GTEx open-access data, including summary statistics of sex-biased genes and sex-biased eQTLs, are available on the GTEx Portal (https://gtexportal.org/home/datasets). Histological images and their annotations are also available on the portal (https://gtexportal.org/home/histologyPage). GTEx v8 sex-stratified eQTL data is available on LocusCompare (http://locuscompare.com/). All GTEx protected data are available via dbGaP (accession phs000424.v8). Access to the raw sequence data is now provided through the AnVIL platform (https://gtexportal.org/home/protectedDataAccess). Code for the sex-biased gene expression analysis is deposited at https://zenodo.org/record/3939042 (doi:10.5281/zenodo.3939042).

Supplementary Materials:

Materials and Methods

Figures S1-S10

Tables S1-S15

References (*73-102*)

Figure legends

Fig. 1. Sample, data types and discovery sets in the study of sex differences in GTEx v8. Illustration of the tissue types (including 11 distinct brain regions and 2 cell lines), with sample numbers from GTEx v8 genotyped donors in parentheses (females:males) and color coding indicated in the adjacent circles. N=44 tissue sources present in both sexes with ≥70 samples were included in this study. Tissue sources comprise two cell lines, 40 tissues and two additional replicates for brain cerebellum and cortex tissues. Tissue name abbreviations are shown in bold. The specific number of donors used in each analysis is stated in the corresponding Supplementary Methods section (*9*).

Fig. 2. Sex-differential gene expression. (**A**) Number of sex-differentially expressed genes (sex-biased genes) per tissue. See Fig. 1 for the legend of tissue colors. (**B**) Sex-biased gene discovery (histogram, number of sex-biased genes) and characteristics of sex-biased genes (stacked bar plots) as a function of tissue-sharing. The characteristics are: *Chr*. for the proportion of X-linked and autosomal sex-biased genes, *Sign* for the proportion of female-biased (F) and male-biased (M) genes. (**C**) Hierarchical clustering of tissues based on gene expression (left) and the effect size of sex-biased genes (right). See (*9*) for further details.

Fig. 3. Regulatory mechanisms and biological functions of sex-biased genes. (**A**) Genomic position enrichment of sex-biased genes; based on male-biased (blue), and female-biased (red) genes across all chromosomes (left) and chromosome X (right). The y-axis represents the tissue-sharing of the significant genomic enrichment signal and ranges from 1 to 44 (number of tissue sources). See (*9*) for further details. (**B**) Transcription Factor Binding Site (TFBS) enrichment in promoter regions of sex-biased genes. Out of 92 enriched TFBS profiles, the top 40 with the largest difference across all tissues in the enrichment profile derived from male-biased and female-biased genes are displayed. Values represent the TFBS enrichment ranking transformed to [0, 1] per tissue and per sex; a value of 1 corresponds to the highest enrichment. See (*9*) for further details. (**C**) Clusters (grey circles) of gene sets enriched for genes highly expressed (blue and red balloons) in females (red) or males (blue) across tissues. Balloon size corresponds to the p-value for the across-tissue meta-analysis of GSEA. Faint lines connecting balloons correspond to shared leading edge genes between gene sets. See (*9*) for further details.

Fig. 4. Sex-biased eQTLs (sb-eQTLs). (**A**) Number of sb-eQTLs discovered per tissue. Square-root transformation was applied to the x-axis. See Fig. 1 for tissue abbreviations. (**B**) Association p-values of the female- (top) and male-stratified (bottom) *cis*-eQTLs in the *ADRA1A* locus in adipose subcutaneous tissue (top panels, β_F = -0.78, P_F = 4.64e-18, β_M = -0.47, P_M = 3.98e-10, $P_{G\times Sex}$ = 1.05e-05) and *C4BPB* locus in breast mammary tissue (lower panels, β_F = 0.40, P_F = 2.68e-07, β_M = -0.02, P_M = 8.90e-01, $P_{G\times Sex}$ = 7.22e-05). Linkage disequilibrium between loci is quantified by squared Pearson coefficient of correlation (r^2). Diamond-shaped point represents the top significant eQTL variant across sex-stratified p-values. (**C**) sb-eQTL mediation analysis of 261 breast sb-eQTLs. Point coordinates represent the effect size of the terms G×Sex (x-axis) and G×Epithelial cells (y-axis) derived from a linear regression model with both interaction terms. Grey lines represent confidence intervals of the effect sizes of G×Sex (horizontal lines) and

G×Epithelial cells (vertical lines) terms. Point size represents sb-eQTL significance and color corresponds to mediation significance. See (*9*) for further details.

Fig. 5. Colocalization of sex-biased eQTLs (sb-eQTLs) with GWAS traits (**A**) Posterior probability (PP4) of 74 colocalized gene-trait pairs where a GWAS shows evidence of colocalization with the female- and/or the male-stratified *cis*-eQTL signal (PP4 > 0.5). Legend: in parentheses, number of colocalizing loci per tissue. (**B**) Number of colocalizing loci for female and male *cis*-eQTLs. (**C**) GWAS-eQTL colocalizing genes (PP4 > 0.5) color-labeled by eQTL tissue of origin according to labels in Fig. 5A (x-axis), are categorized by the sex where the colocalization signal is maximized with the corresponding GWAS trait (y-axis). Comparing the colocalization PP4 values for male and female *cis*-eQTL signal, the estimates can be maximum in females (red) or males (blue). (**D**) Genotype-phenotype association p-values of the *CCDC88C* (left panel) and *HKDC1* (right panel) loci. For the *CCDC88C* locus, panels illustrate GWAS signal for breast cancer (top) and *CCDC88C cis*-eQTL signal for females (middle) and males (bottom) in breast mammary tissue. For the *HKDC1* locus, panels illustrate GWAS signal for birth weight (top) and *HKDC1 cis*-eQTL signal for females (middle) and males (bottom) in liver. (**E**) Genotype-phenotype association p-values of the *CLDN7* (left panel) and *DPYSL4* (right panel) loci. For the *CLDN7* locus, panels illustrate GWAS signal for birth weight (top) and *CLDN7 cis*-eQTL signal for females (middle) and males (bottom) in breast mammary tissue. For the *DPYSL4* locus, panels illustrate GWAS signal for body fat (top) and *DPYSL4 cis*-eQTL signal for females (middle) and males (bottom) in muscle skeletal tissue. In (**D**) and (**E**), linkage disequilibrium between loci is quantified by squared Pearson coefficient of correlation (r^2). Diamond-shaped point represents the top significant *cis*-eQTL variant across sex-stratified p-values.

Note:

GTEx Consortium author information has been omitted for brevity. See the published version of the manuscript for the full list of consortium members and their affiliations.

Figure 1

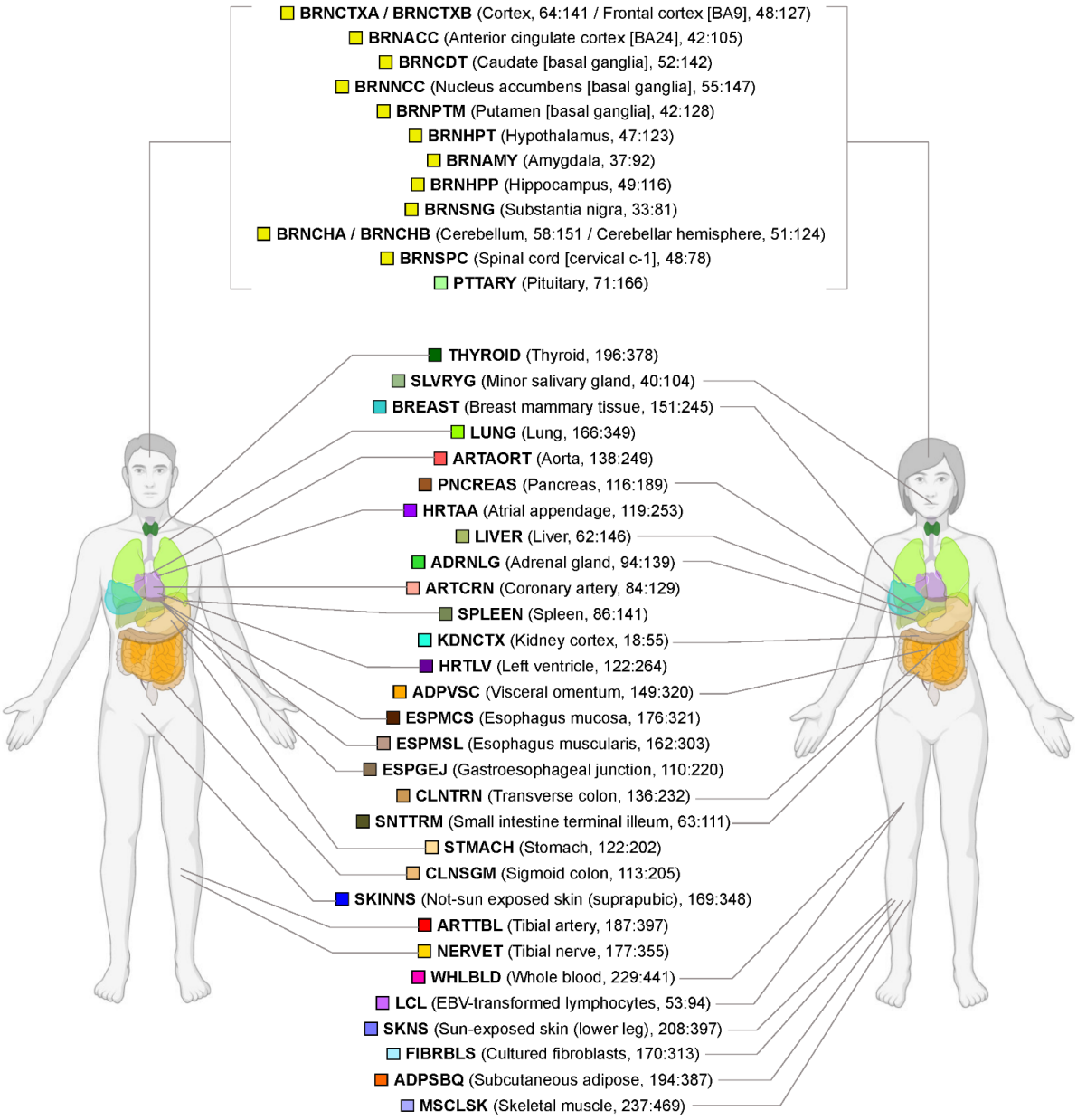


Figure 2

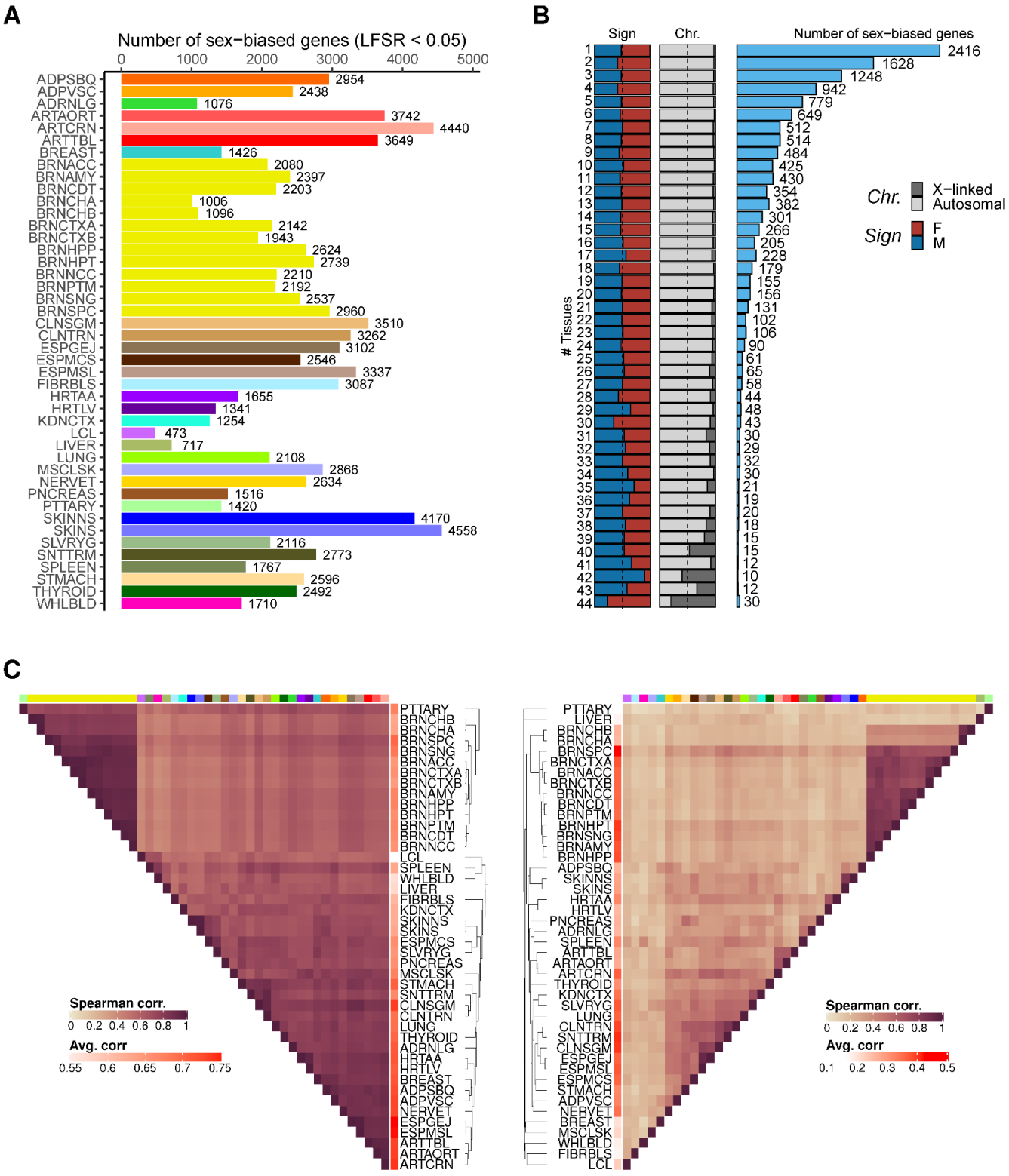


Figure 3

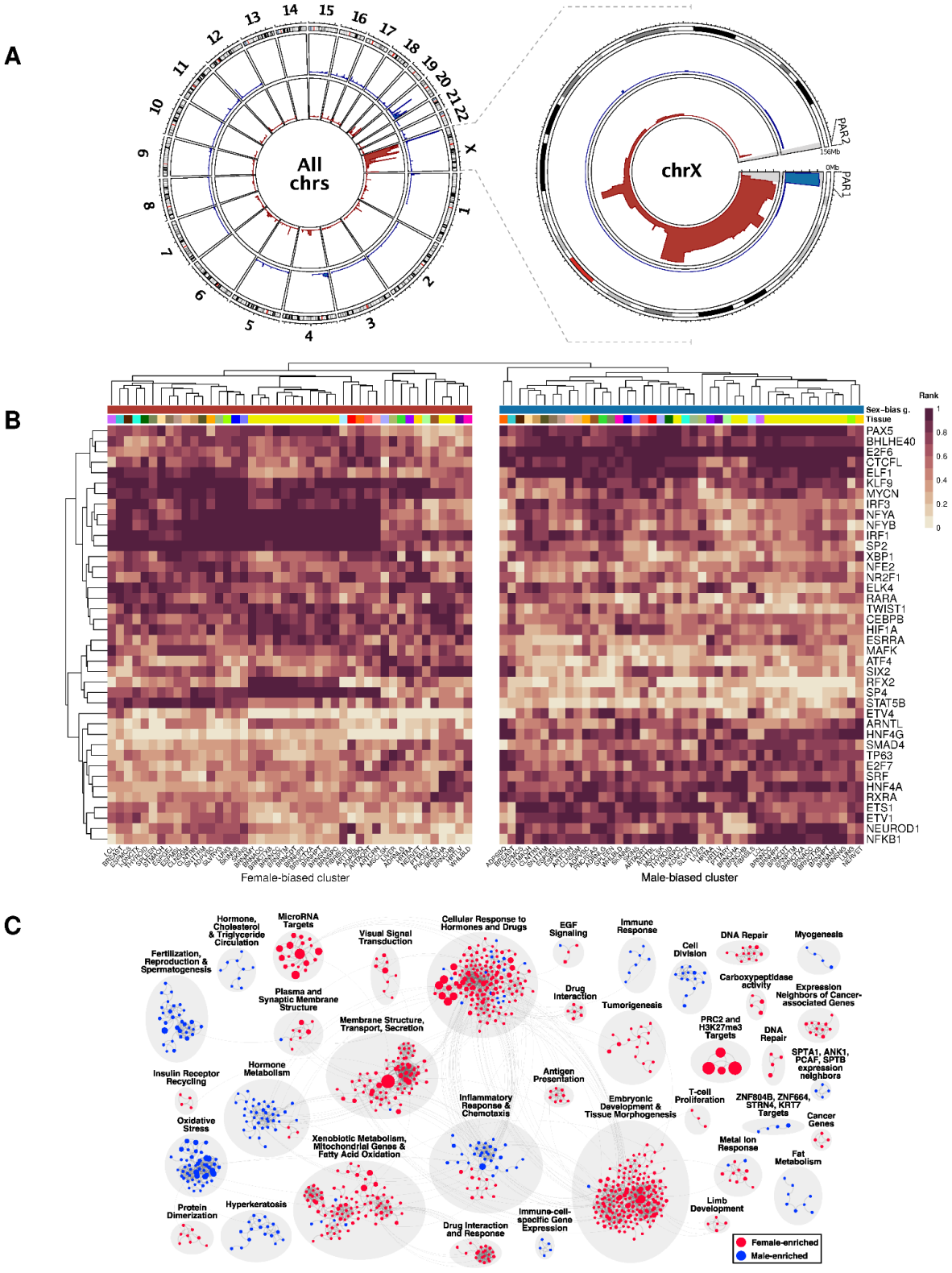


Figure 4

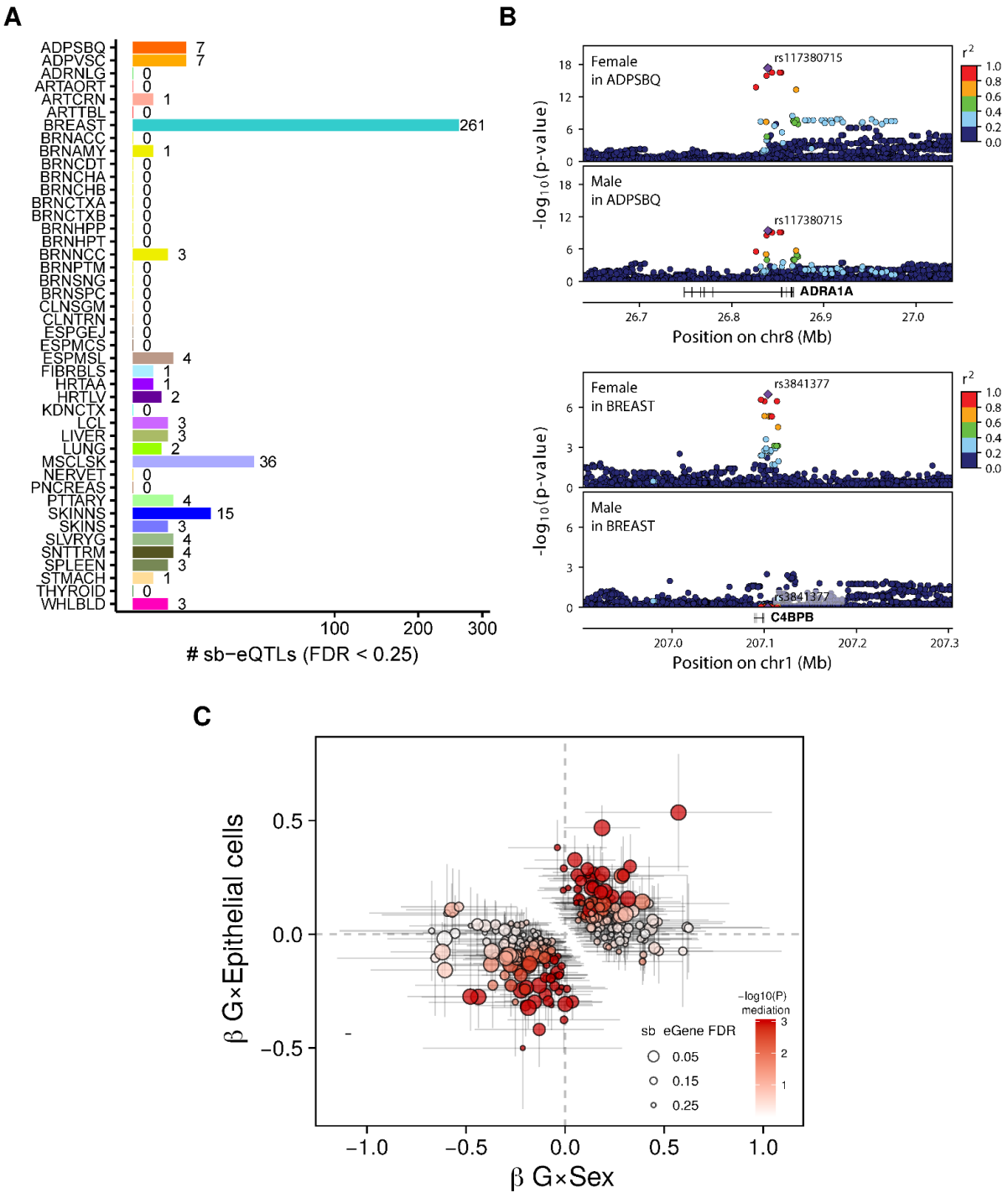
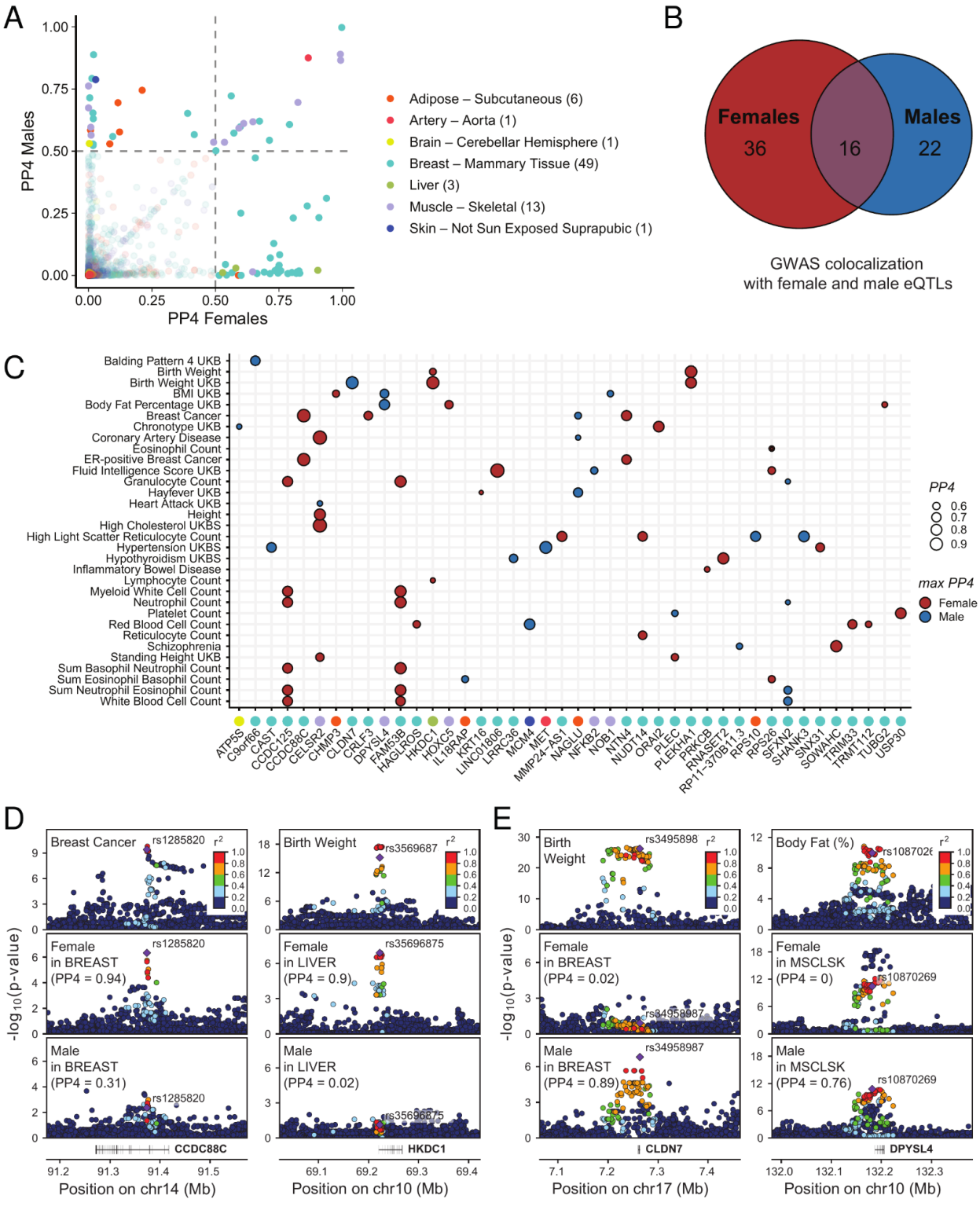


Figure 5



THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

In-silico spatial transcriptomics

In Section 1.1.4 we discussed the importance of characterizing gene expression at the cell type level. However, there is yet another critical component when trying to understand the underlying mechanisms of biological systems: spatial information. The physical location of cells plays a role in gene expression, for example, in the case of tumours where the cells constituting it can differ with respect to the histological patterns of the tissue, and these changes been shown to correlate with alterations at the molecular level [167]. This also means that characterizing spatial patterns can lead to the identification of cell type markers, allowing to refine existing descriptions of the molecular organization of organs and tissues, such as the one proposed in Chapter 4. Several technologies have been developed to disentangle the spatial organization of molecular features, and can be divided into five categories, which we briefly enlist and summarize (based on the descriptions in [168]):

1. *microdissected gene expression* (laser capture microdissection, NICHE-Seq, ProximID): gene expression profiles are obtained from small isolated regions of a sample, presenting challenges to understand spatial organization across a complete tissue but also allowing to analyze isoform presence.

2. *in situ hybridization technologies* (smFISH, seqFISH, MERFISH): RNA molecules are visualized directly in the environment through labeled probe hybridization, allowing to detect a large set of transcripts but suffering from spectral overlap.

3. *in situ sequencing technologies* (ISS with padlock probes, BaristaSeq, STARmap, FISSEQ): sequencing is directly performed on the RNA of a cell within the tissue, using DNA balls to amplify the signal allowing to resolve the subcellular location of the transcript but with limits for transcript discrimination.

4. *in situ capturing technologies* (spatial transcriptomics/10x Visium, Slide-Seq, APEX-Seq, microfluidic barcoding): transcripts are captured in situ but sequenced ex situ, allowing unbiased analyses of the whole transcriptome, but subject to RNA capture inefficiencies. The spatial transcriptomics technique was first introduced by Ståhl et al. [169], where they performed RNA-seq mantaining 2D-positional information in the tissue through arrayed reverse transcription primers with unique positional barcodes, over mouse brain and human breast cancer tissues. See Fig. 6-1 for an example.

5. *in silico reconstruction of spatial data*: computational approaches used to assign a spatial location to dissociated single cells based on their expression profiles, either using a reference map (gene signatures) or *de novo*, through assumptions of how the gene expression will vary across the tissue and modeled through optimization problems that seek to preserve distances both with respect to gene expression feature space but also with respect to the spatial organization.

Spatial localization of gene expression has already proven useful in the context of disease: for example, Maniatis et al. [170] have examined mouse spinal cords and postmortem human tissue affected by amyotrophic lateral sclerosis (ALS), using the spatial transcriptomics technique to spatiotemporally characterize gene expression, identifying shared transcriptional pathways perturbated in the context of ALS in both human and mouse, as well as spatial differences in cell type populations.

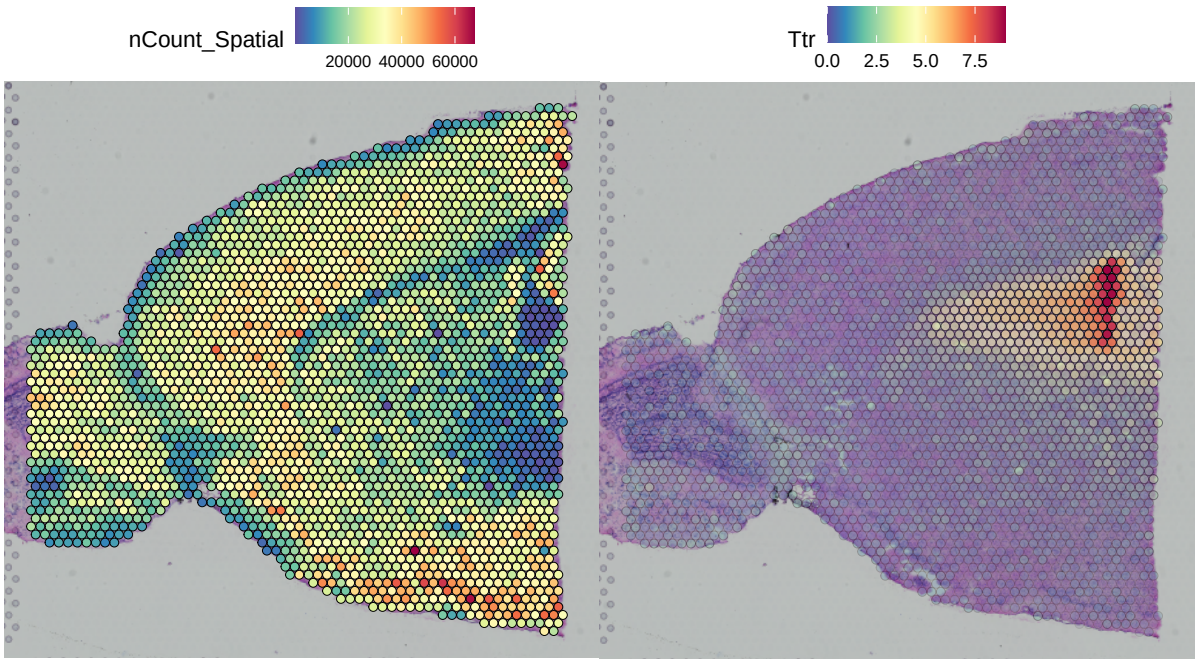


Figure 6-1: **Visium (10x Genomics) spatial transcriptomics**. Sagital mouse brain slice generated with Visium v1 chemistry. The left plot shows the total number of molecular counts, while the right plot shows the normalized expression for *Ttr*, a marker of the choroid plexus. Data obtained from [171], with figures reproduced as described in the Seurat R package vignette [172].

6.1 A framework for in-silico spatial transcriptomics

In this chapter, we aim to design and implement a framework to spatially resolve gene expression, similarly to the spatial transcriptomics protocols described before, but in an in-silico (data driven) manner, with matched sample pairs of histological images and molecular traits such as bulk RNA-seq gene expression. In other words, the goal is to implicitly verify if it is possible to associate patterns of gene expression variation with histological patterns encoded in the WSIs of the corresponding tissue samples, at a level of resolution that is reasonable enough to identify subregions of tissues that are known to be associated with specific patterns of gene expression. To this end, we will use the GTEx data resource since it is currently one of the largest datasets that satisfy the matched data pairs assumption.

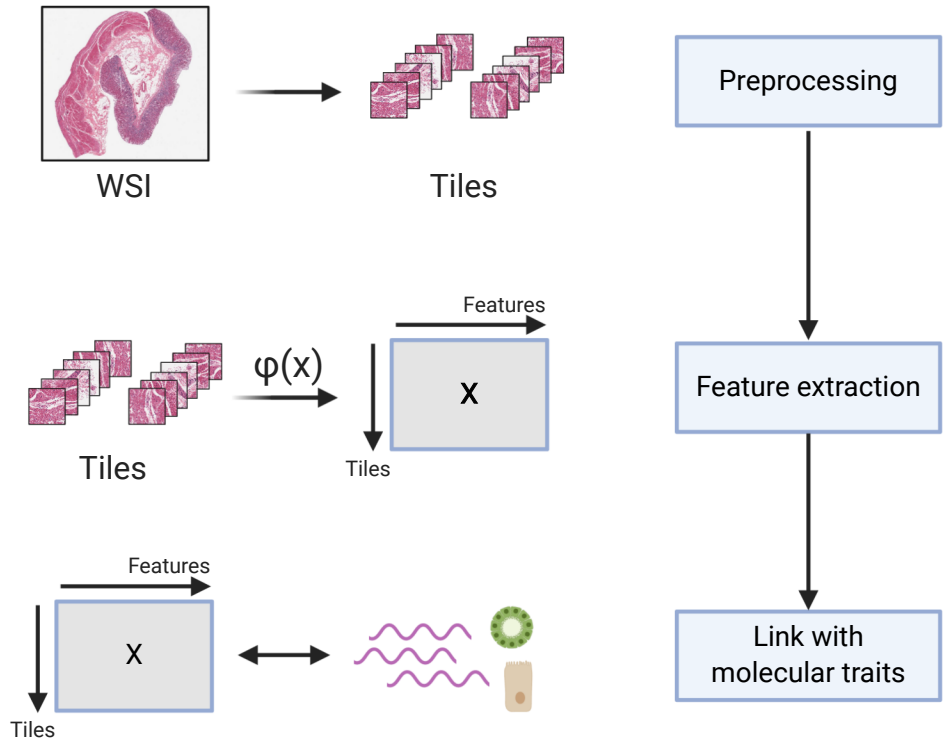


Figure 6-2: **In-silico spatial transcriptomics pipeline**. In the preprocessing steps, the WSIs corresponding to samples from many tissues will be divided into square tiles. The second step consists in mapping the pixel representation of a tile into a feature vector of reduced dimensionality that encodes the relevant histological features of the tile. The third step is the association of the matrix of tile features (or their summarization) with molecular traits.

Although here we do not have experimental data of spatially-resolved gene expression, our working hypothesis is that if the gene expression patterns (even if at a bulk level) are consistent with variation in histological primitives, it might be possible to pinpoint the expression of specific genes (at least, those with known tissue specificity and relatively high levels of expression) in the WSIs. Without considering specific implementation details for now, such a framework might consist of at least the following three general steps, as illustrated in Fig. 6-2. The first one corresponds to the preprocessing of the WSIs. We have introduced in Section 1.1.6 the generalities of how WSIs are constituted and how their large size poses constraints in deciding how any computation over them will be performed, since, at least in a GPU-based computational model, due to hardware limitations it is currently not possible to fit many WSIs entirely in memory (since memory allocation occurs not only for data but also for model parameters).

In the second step, we aim to learn a function $\phi(\cdot)$ that maps a tile $\mathbf{P} \in \mathbb{R}^{n \times p \times 3}$ (where n and p correspond to the tile width and height across three color channels, and in practice, with $n = p$) into a compressed representation in the form of feature vector $x \in \mathbb{R}^q$ that would ideally capture the histological patterns encoded in the image tile.

Lastly, in the third step we use these compressed representations to characterize associations with molecular traits (such as gene expression). This last step poses several questions, such as, for example, determining a way to aggregate tile representations from a single sample into a summarization at the WSI-level, or how to choose an initial set of genes to validate the framework. The specific details for these steps will be discussed in the next sections, together with a discussion of the computational experiments performed so far, since this is still a work in progress.

6.2 Image preprocessing

Before delving into the problem of linking gene expression variation with patterns in histological images, a practical consideration needs to be addressed: the preprocessing of the WSIs. Since these are large in terms of size (pixels, and also in storage space), for most types of histological image analysis techniques it is unfeasible to process them at once as a single unit. Thus, it is common procedure in the histological image analysis literature to break down the image into square tiles and perform the relevant modeling at that level. Since this also holds true for the particular methodology to attempt in-silico spatial transcriptomics, this section will describe considerations in the problem of extracting tiles from histological images. The process of extracting tiles from a WSI is not standardized and researchers often need to develop custom software to perform this task whenever building a model. In the problem we explore in this chapter, we do not need to consider background tiles. Discarding these is desirable since background tiles do not add anything to the model: associations between gene expression and the background should not exist, and even if we included them when training a model and it correctly learned to recognize that these tiles do not contribute anything, we would still be incurring in unnecessary computational overhead.

6.2.1 Tile extraction generalities

To this end, as an initial experiment, we propose to frame tile extraction as a classification problem at the tile level: we aim to keep foreground tiles (positive class) and discard the background tiles (negative class). First, we define a whole slide image $\mathbf{I}$ and a patch $\mathbf{P}$ as:

$$\mathbf{P}_{ij} \in \mathbb{R}^{n \times p \times 3};\ i \in \{1, \ldots, a\}, j \in \{1, \ldots, b\} \tag{6.1}$$

$$\mathbf{I} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \ldots & \mathbf{P}_{1b} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \ldots & \mathbf{P}_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{a1} & \mathbf{P}_{a2} & \ldots & \mathbf{P}_{ab} \end{bmatrix} \in \mathbb{R}^{(n \times a) \times (b \times p) \times 3} \tag{6.2}$$

where a refers to the number of tiles in the height axis, b to the number of tiles in the width axis, and n and p to the number of pixels in the respective axes across the three color channels.

To generate a response vector $\mathbf{y}$ with the ground truth labels for tiles, we sort the tiles by increasing file size, and keep the top k and bottom k tiles, assigning them positive and negative labels, respectively. The rationale behind this is that at the top and bottom of the ranking, tiles are likely to have large amounts of tissue content while the ones at the bottom will be mostly empty, as illustrated in Fig. 6-3. This works since more complex images need to encode more information and are thus larger in file size (at least when using compressed image file formats).

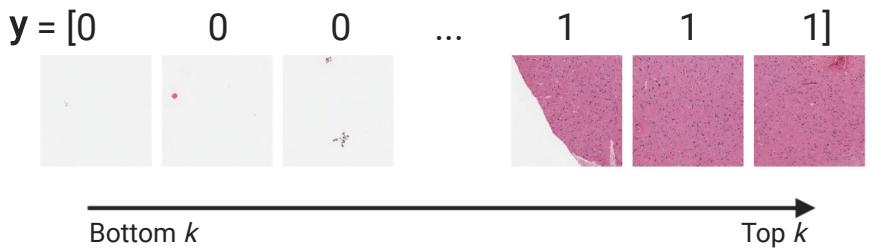


Figure 6-3: **Tile label assignment**. Examples of tiles sorted by their increasing file size. The top and bottom k tiles are chosen to be part of the dataset, assigning them positive and negative labels, respectively.

We generate a feature matrix $\mathbf{X}$ to use in a classifier model by vectorizing each patch into a row vector and stacking them into a matrix, in such a way that $\mathbf{X} = \mathbf{I} \rightarrow \mathbb{R}^{(a \times b) \times (n \times p \times 3)}$. Then, PCA can be performed over $\mathbf{X}$ to obtain principal components $\mathbf{T}$ out of which a limited number can be used as a final set of features. Note that this process has to be performed independently for tiles assigned into training, testing and validation sets in order to avoid data leakage. Then, any machine learning algorithm, for example a random forest, can be used to predict a response vector $\hat{\mathbf{y}}$ for these tiles indicating if they should be kept or not.

Although this approach is quite naive and works reasonably well for many different types of tissue, it has two main drawbacks: i) as discussed in Section 2.1.1 the spatial relationships are lost, and ii) it can prove troublesome to use for slides where the background could easily be mistaken for part of the tissue, such as for example,

adipose tissue (see Fig. C-10). Thus, modeling tile extraction as a classification problem would require careful labeling of the training data for the tiles, as well as using a more sophisticated approach (for example, a neural network) to perform feature extraction from the tiles in order to classify them.

To propose a more robust and lightweight solution for the tile extraction problem, and to provide flexibility in the way that image tiles can be generated, we developed a software solution: *PyHIST: A Histological Image Segmentation Tool*, which is covered in the article that follows.

Summary of my key contributions

- A command-line based Python tool to extract custom-sized square tiles from WSIs in SVS/TIFF/NDPI/VMS formats. The native resolution of the WSI can be used to perform the extraction, but resizing operations to other resolutions (in powers of 2) are also supported.

- Mask generation for the tissue foreground: to identify tissue slices in the WSI to retrieve the tiles, a mask for the WSI is generated to discriminate the background from the foreground. Three methods are available to generate the mask: graph-based segmentation, adaptive thresholding, and Otsu thresholding. Tiles are then extracted from the foreground.

- Segmentation overview: The tool generates an overview image to quickly identify how the tiling grid appears when overlaid on top of the WSI, as well as indicators of which tiles were selected as foreground.

- Random tile sampling: For those applications that do not need to distinguish foreground from background, the tool also provides the possibility to sample tiles from random starting positions in the WSI.

- Containerization: All the dependencies to run the tool are bundled in a Docker image to provide cross-platform support (Windows/macOS/Linux).

- An use-case demonstrating the applicability of the tool: we create a classifier model using transfer learning with a ResNet-152 convolutional neural network to identify the tissue of origin for 7163 tiles extracted from WSIs of six cancer-affected tissues from The Cancer Genome Atlas (glioblastoma in brain, infiltrating duct carcinoma in breast, adenocarcinoma in colon, clear cell carcinoma in kidney, hepatocellular carcinoma in liver, and malignant melanoma in skin). We demonstrate that the generated feature vectors for the tiles recapitulate tissue morphology by observing the presence of six tissue clusters. This use case is documented in three reproducible Jupyter notebooks.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

PyHIST: A Histological Image Segmentation Tool

Manuel Muñoz-Aguirre[1,2☯]*, **Vasilis F. Ntasis**[1☯], **Santiago Rojas**[3], **Roderic Guigó**[1,4]

1 Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain, **2** Department of Statistics and Operations Research, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain, **3** Unit of Human Anatomy and Embryology. Department of Morphological Sciences. Faculty of Medicine. Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Catalonia, Spain, **4** Department of Experimental and Health Sciences (DCEXS), Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

☯ These authors contributed equally to this work.
* manuel.munoz@crg.eu

Citation: Muñoz-Aguirre M, Ntasis VF, Rojas S, Guigó R (2020) PyHIST: A Histological Image Segmentation Tool. PLoS Comput Biol 16(10): e1008349. https://doi.org/10.1371/journal.pcbi.1008349

Editor: Dina Schneidman-Duhovny, Hebrew University of Jerusalem, ISRAEL

Received: May 20, 2020

Accepted: September 17, 2020

Published: October 19, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008349

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The authors received no specific funding for this work. M.M.-A. performs his research with

Abstract

The development of increasingly sophisticated methods to acquire high-resolution images has led to the generation of large collections of biomedical imaging data, including images of tissues and organs. Many of the current machine learning methods that aim to extract biological knowledge from histopathological images require several data preprocessing stages, creating an overhead before the proper analysis. Here we present PyHIST (https://github.com/manuel-munoz-aguirre/PyHIST), an easy-to-use, open source whole slide histological image tissue segmentation and preprocessing command-line tool aimed at tile generation for machine learning applications. From a given input image, the PyHIST pipeline i) optionally rescales the image to a different resolution, ii) produces a mask for the input image which separates the background from the tissue, and iii) generates individual image tiles with tissue content.

Author summary

Histopathology images are an essential tool to assess and quantify tissue composition and its relationship to disease. The digitization of slides and the decreasing costs of computation and data storage have fueled the development of new computational methods, especially in the field of machine learning. These methods seek to make use of the histopathological patterns encoded in these slides with the aim of aiding clinicians in healthcare decision-making, as well as researchers in tissue biology. However, in order to prepare digital slides for usage in machine learning applications, researchers usually need to develop custom scripts from scratch in order to reshape the image data in a way that is suitable to train a model, slowing down the development process. With PyHIST, we provide a toolbox for researchers that work in the intersection of machine learning, biology and histology to effortlessly preprocess whole slide images into image tiles in a standardized manner for usage in external applications.

support of pre-doctoral fellowship FPU15/03635 from Ministerio de Educación, Cultura y Deporte. (URL: http://www.mecd.gob.es/) Agencia Estatal de Investigación (AEI) and FEDER under project PGC2018-094017-B-I00 is also acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

This is a *PLOS Computational Biology* Software paper.

Introduction

In histopathology, Whole Slide Images (WSI) are high-resolution images of tissue sections obtained by scanning conventional glass slides [1]. Currently, these glass slides of fixed tissue samples are the preferred method in pathology laboratories around the world to make clinical diagnoses [2], notably in cancer [3]. However, the increasing automation of WSI acquisition has led to the development of computational methods to process the images with the goal of helping clinicians and pathologists in diagnosis and disease classification [4]. As an increasing number of larger WSI datasets became available, methods have been developed for a wide array of tasks, such as the classification of breast cancer metastases, Gleason scoring for prostate cancer, tumor segmentation, nuclei detection and segmentation, bladder cancer diagnosis, mutated gene prediction, among others [5–10]. Besides of being important diagnostic tools, histopathological images capture endophenotypes (of organs and tissues) that, when correlated with molecular and cellular data on the one hand, and higher-order phenotypic traits on the other, can provide crucial information on the biological pathways that mediate between the sequence of the genome and the biological traits of the organisms (including diseases) [11].

Because of the complexity of the information typically contained in WSIs, Machine Learning (ML) methods that can infer, without prior assumptions, the relevant features that they encode are becoming the preferred analytical tools [12]. These features may be clinically relevant but challenging to spot even for expert pathologists, and thus, ML methods can prove valuable in healthcare decision-making [13].

In most ML tasks, data preprocessing remains a fundamental step. Indeed, in the domain of histological images, there are several issues when preprocessing the data before an analysis: due to the large dimensions of WSIs, many deep learning applications have to break them down into smaller-sized square pieces called tiles [14]. Furthermore, a significant fraction of the area in a WSI is often uninformative background that is not meaningful for the majority of downstream analyses. To circumvent this, some applications apply a series of image transformations to identify the foreground from the background (see, for example, [15]), and perform relevant operations only over regions with tissue content. However, this process is not standardized, and customized scripts have to be frequently developed to deal with data preparation stages (see, for example [10,15]). This is cumbersome and may introduce dataset specific-biases, which can prevent integration across multiple datasets.

Currently available tools for WSI processing focus mostly on the analysis of human-interpretable features by means of nuclei segmentation, object quantification and region-of-interest annotation [16–18]; but WSI preparation into tiles for external ML applications has not yet been directly addressed. To systematize the WSI preprocessing procedure for these applications, and in order to streamline the data preparation stage at the initial phase of a ML project by avoiding the need of creating custom image preprocessing scripts, we developed PyHIST, a command-line based pipeline to segment the regions of a histological image into tiles with relevant tissue content (foreground) with little user intervention. PyHIST was developed to process Aperio SVS/TIFF WSIs due to this format being supported by large slide databases such as The Cancer Genome Atlas (TCGA) which has approximately 31,000 WSIs [19] and The Genotype-Tissue Expression Project (GTEx) with approximately 25,000 WSIs [20]. PyHIST currently has experimental support for other image formats (see S1 Text).

Design and implementation

PyHIST is a command-line Python tool based on OpenSlide [21], a library to read high-resolution histological images in a memory-efficient way. PyHIST's input is a WSI encoded in SVS format (Fig 1A), and the main output is a series of image tiles retrieved from regions with tissue content (Fig 1E).

The PyHIST pipeline involves three main steps: 1) produce a mask for the input WSI that differentiates the tissue from the background, 2) create a grid of tiles on top of the mask, evaluate each tile to see if it meets the minimum content threshold to be considered as foreground and 3) extract the selected tiles from the input WSI at the requested resolution. By default, PyHIST uses a graph-based segmentation method to produce the mask. In this method, first, tissue edges inside the WSI are identified using a Canny edge detector (Fig 1B), generating an alternative version of the image with diminished noise and an enhanced distinction between the background and the tissue foreground. Second, these edges are processed by a graph-based segmentation algorithm [22], which is used here to identify tissue content. In short, this step evaluates the boundaries between different regions of an image as defined by the edges;

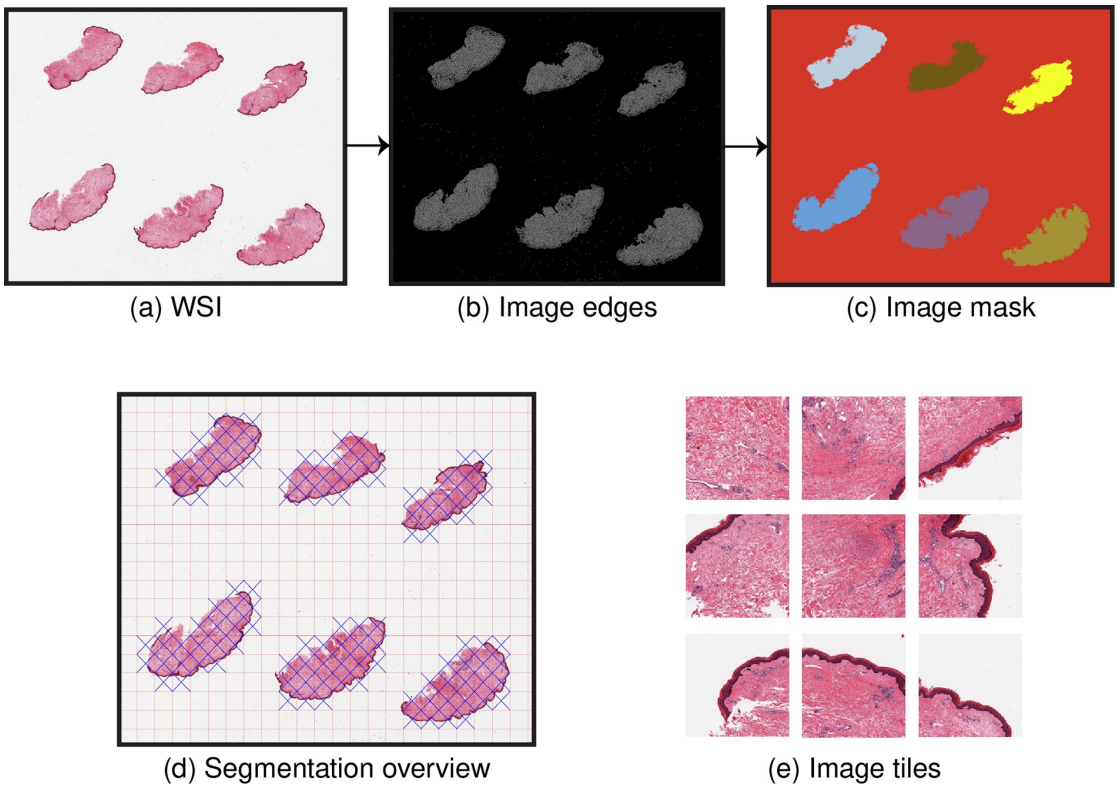


Fig 1. PyHIST pipeline. (a) The input to the pipeline is a Whole Slide Image (WSI). Within PyHIST, the user can decide to scale down the image to perform the segmentation and tile extraction at lower resolutions. The WSI shown is of a skin tissue sample (GTEX-1117F-0126) from the Genotype-Tissue Expression (GTEx) project [20]. (b) An alternative version of the input image is generated, where the tissue edges are highlighted using a Canny edge detector. A graph segmentation algorithm is employed over this image in order to generate the mask shown in (c). PyHIST extracts tiles of specific dimensions from the masked regions, and provides an overview image to inspect the output of the segmentation and masking procedure, as shown in (d), where the red lines indicate the grid generated by tiling the image at user-specified tile dimensions, while the blue crosses indicate the selected tiles meeting a certain user-specified threshold of tissue content with respect to the total area of the tile. In (e), examples of selected tiles are shown.

https://doi.org/10.1371/journal.pcbi.1008349.g001

different parts of the image are represented as connected components of a graph, and the "within" and "in-between" variations of neighboring components are assessed in order to decide if the examined image regions should be merged or not into a single component. From this, a mask is obtained in which the background and the different tissue slices are separated and marked as distinct objects using different colors (Fig 1C). Finally, the mask is divided into a tile grid with a user-specified tile size. These tiles are then assessed to see if they meet a minimum foreground (tissue) threshold with respect to the total area of the tile, in which case they are kept, and otherwise are discarded. Optionally, the user can also decide to save all the tiles in the image.

Of note, tile generation can be performed at the native resolution of the WSI, but downsampling factors can also be specified to generate tiles at lower resolutions. Additionally, edge detection and mask generation can also be performed on downsampled versions of WSIs—reducing segmentation runtimes (S1 Fig, S1 Text). A segmentation overview image is generated at the end of the segmentation procedure for the user to visually inspect the selected tiles (Fig 1D). With the set of parameters available in PyHIST (S2 Text), the user can specify regions to ignore when performing the masking and segmentation (S2 Fig), and have a fine-grained control over specific use-cases.

By default, PyHIST uses the graph-based segmentation method described previously due to its robustness in detecting tissue foreground in WSIs that do not have a homogeneous composition. However, alternative tile-generation methods based on thresholding that tend to work well on heterogeneous WSIs are also implemented (S3–S5 Figs, see S1 Text for details and benchmarking information). PyHIST also has a random tile sampling mode for those applications that do not necessarily need to distinguish the background from the foreground. In this mode, tiles at a user-specified size and resolution will be extracted from random starting positions in the WSI.

Results

To demonstrate how PyHIST can be used to preprocess WSIs for usage in a ML application, we generated a use case example with the goal of building a classifier at the tile-level that allows us to determine the cancer-affected tissue of origin based on the histological patterns encoded in these tiles. To this end, we first retrieved a total of 36 publicly available WSIs, six from each of the following human tissues hosted in The Cancer Genome Atlas (TCGA) [23]: Brain (glioblastoma), Breast (infiltrating ductal carcinoma), Colon (adenocarcinoma), Kidney (clear cell carcinoma), Liver (hepatocellular carcinoma), and Skin (malignant melanoma). Slides within each tissue have the same cancer primary diagnosis as established by TCGA. Second, these WSIs were preprocessed with PyHIST, generating a total of 7163 tiles with dimensions 512x512. These tiles were then partitioned into training and test sets (constraining all the tiles of a given WSI to be in only one of the two sets), and we then fit a deep learning convolutional neural network model over these tiles with weighted sampling at training time (S6 Fig), achieving a classification accuracy of 95% (Fig 2A, S1 Table, S2 Table, see S3 Text for data preparation and model details, and a detailed assessment of Fig 2A).

We also inspected the feature vectors generated by the deep learning model: for each tile, we retrieved the features corresponding to the linear layer of the last (fully connected) sequential container of the model, and performed dimensionality reduction (t-SNE) over the stacked matrix of these vectors. From here, we infer that the learned features recapitulate tissue morphology since tile clusters corresponding to each tissue are formed (Fig 2B, S7 Fig). We note that this classifier is only an exercise to show end-users how to quickly prepare WSI data using PyHIST to generate tiles, reducing the overhead to start performing downstream analyses:

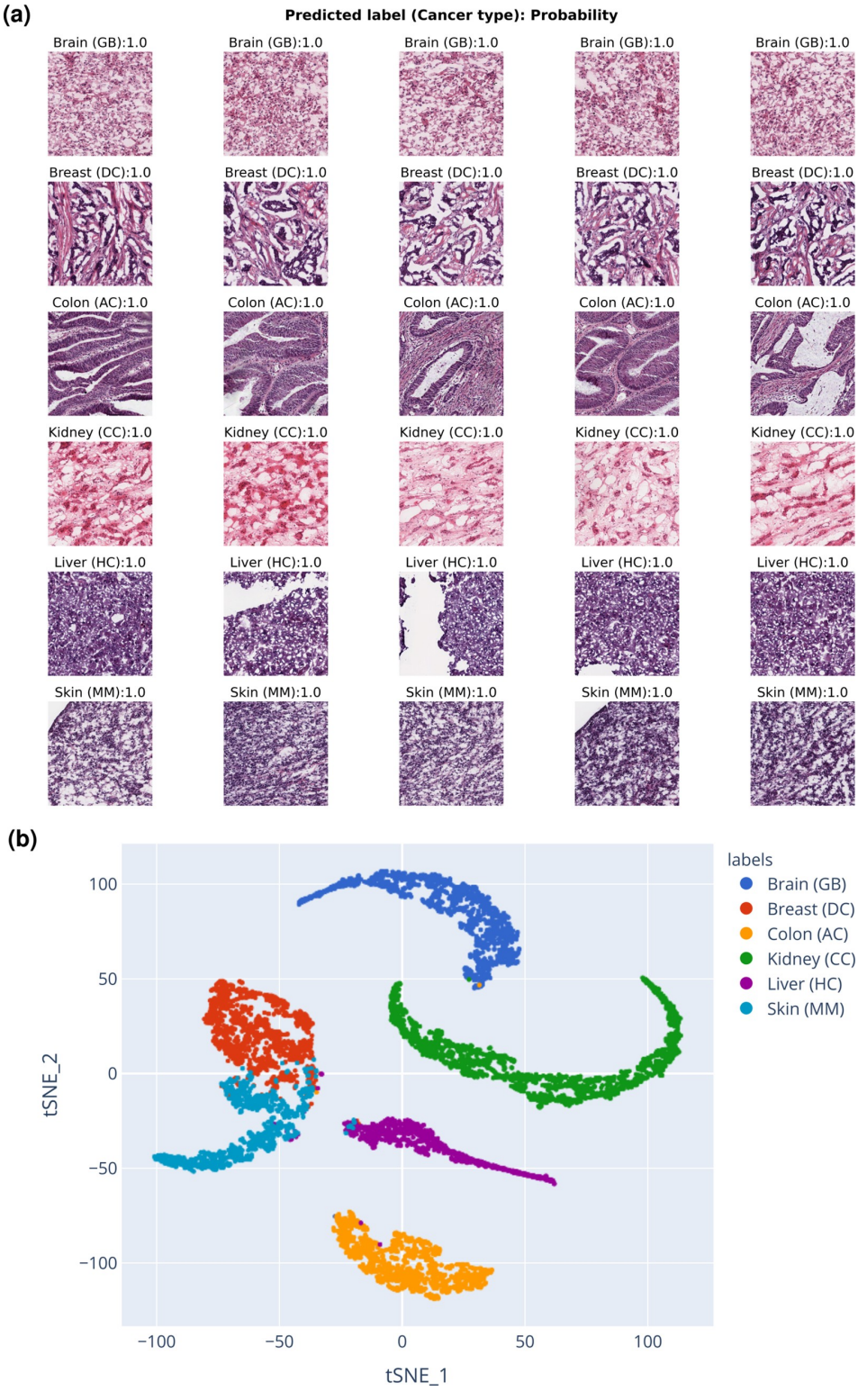

(a)
Predicted label (Cancer type): Probability
Brain (GB):1.0
Breast (DC):1.0
Colon (AC):1.0
Kidney (CC):1.0
Liver (HC):1.0
Skin (MM):1.0
(b)
labels
Brain (GB)
Breast (DC)
Colon (AC)
Kidney (CC)
Liver (HC)
Skin (MM)
tSNE_1
tSNE_2
100
50
0
−50
−100

Fig 2. TCGA use case. (a) Examples of the top 5 most accurately predicted tiles per cancer-affected tissue (rows) from the TCGA use case test set. The label above each tile shows the predicted cancer-affected tissue type (GB: glioblastoma, DC: infiltrating ductal carcinoma, AC: adenocarcinoma, CC: clear cell carcinoma, HC: hepatocellular carcinoma, MM: malignant melanoma), followed by the probability of the ground truth label. All of these tiles were correctly classified. (b) Dimensionality reduction of TCGA tiles. t-SNE performed with the feature vectors of each tile that were derived from the deep learning classifier model. Each dot corresponds to an image tile.

https://doi.org/10.1371/journal.pcbi.1008349.g002

further tuning of the model with more data is desirable to ensure that the classifier is robust enough to generalize to different types of unseen WSIs for a real application.

Availability and future directions

The example use case described above is documented and fully available at https://pyhist.readthedocs.io/en/latest/testcase/, and divided into three Jupyter notebooks: 1) Data preprocessing with PyHIST, 2) Constructing a deep learning tissue classifier, and 3) Dimensionality reduction. The TCGA WSIs in the use case were downloaded from the Genomic Data Commons (GDC) repository (https://gdc.cancer.gov/) using the GDC Data-transfer tool (https://gdc.cancer.gov/access-data/gdc-data-transfer-tool).

PyHIST is a generic tool to segment histological images automatically: it allows for easy and rapid WSI cleaning and preprocessing with minimal effort to generate image tiles geared towards usage in ML analyses. The tool is available at https://github.com/manuel-munoz-aguirre/PyHIST and released under a GPL license. Updated documentation and a tutorial can be found at https://pyhist.readthedocs.io/. PyHIST is highly customizable, enabling the user to tune the segmentation process in order to suit the needs of any particular application that relies on histological image tiles. The software and all of its dependencies have been packaged in a Docker image, ensuring portability across different systems. PyHIST can also be used locally within a regular computing environment with minimal requirements. Future directions and improvements include adding support for more histological image formats and features to save tiles into specialized data structures, as well as the inclusion of a graphical user interface to ease the learning curve for users who are new to the field of image processing for ML analyses. Finally, PyHIST is open source software: all the code and reproducible notebooks for the example use case are available in GitHub and will continue to be improved based on user feedback.

Supporting information

S1 Text. PyHIST overview. General description of the pipeline: supported file formats, tile generation methods, and execution times.
(PDF)

S2 Text. Parameter description. Description of supported arguments in PyHIST.
(PDF)

S3 Text. TCGA tissue classification use case. Description of data preprocessing, model training and analysis for the TCGA tissue classification use case.
(PDF)

S1 Fig. WSI scaling steps in PyHIST. (a) WSI at its original resolution (1x). (b) The mask can be generated and processed at a given downsampling factor. A smaller resolution will lead to a faster segmentation. (c) The output can be requested at a given downsampling factor. (d) The segmentation overview image can also be generated at a given downsampling factor. The dimensions in all steps are matched to ensure that the tile sizes and grid are consistent. The

downsampling choices for all the steps are independent of each other.
(PNG)

S2 Fig. Image in graph-based segmentation test mode. Test mode allows the user to see how the image mask will be with the chosen segmentation parameters and tile dimension configuration, before proceeding to generate the individual tile files. The black border defines the region of exclusion for tissue content placed within the edges of the slide (see—*borders* and—*corners* arguments, and section 2.2 in S2 Text).
(PNG)

S3 Fig. Comparison of mask generation methods. (a) Adipose tissue WSI from the GTEx project, from sample GTEX-111CU-1826. Thresholding-based masks (b-d) are generated by first converting (a) into grayscale and then applying the corresponding thresholding method. Note that simple thresholding is shown here for completeness but only Otsu and adaptive are implemented in PyHIST due to their overall better performance when compared to simple thresholding. In the graph-based method, an image with highlighted edges is first generated through a Canny edge detector (e, left) and then the connected components are labeled through graph-based segmentation (e, right).
(PNG)

S4 Fig. Runtime benchmarks for random sampling and graph-based segmentation. (a) Execution time to perform random sampling (y-axis) of a varying number of tiles (x-axis) at different downsampling factors for the WSI shown in S1 Fig. For each combination of number of tiles and downsampling factor, the sampling was repeated 30 times. Each dot represents the average running time across the 30 runs, while the interval shows the range between the maximal and minimal running time. (b) Execution time to perform random sampling of 1000 tiles (y-axis) at different tile dimensions (x-axis) at different downsampling factors for the same WSI in (a). Each combination was repeated 50 times, with each dot showing the average runtime. (c) Segmentation runtime of 50 Stomach WSIs from the GTEx project, at different downsampling factors, at a tile size of 256x256. Each dot represents the average execution time. Each interval shows the range between the fastest and slowest segmentations, while the labels show the dimensions of the corresponding WSIs. (d) Segmentation runtime (y-axis) at 1x resolution for the 50 Stomach WSIs, with respect to the number of pixels in the WSI (x-axis).
(PNG)

S5 Fig. Runtime comparison of mask-generating methods. Tile extraction was evaluated for the three different methods at four different settings of tile size. Each method + tile size combination was repeated ten times to show runtime variability.
(PNG)

S6 Fig. Tile distribution per class in a training epoch in the TCGA example use case. Within each training epoch, weighted random sampling is performed to create batches with a fair distribution of tiles among the classes. Even if the sample sizes in the training dataset are different among the classes, the balance in the number of tiles per epoch is obtained through data augmentation.
(PNG)

S7 Fig. Correlation matrix of TCGA tiles based on their feature vectors. Heatmap of Pearson's correlation matrix between the feature vectors obtained for each TCGA tile. Rows and columns are reordered with hierarchical agglomerative clustering.
(PNG)

S1 Table. Tile distribution across classes in the TCGA use case training and test sets.
(PNG)

S2 Table. Confusion matrix for the tiles in the test set of the TCGA use case.
(PNG)

Acknowledgments

We acknowledge Kaiser Co and Valentin Wucher for testing PyHIST, and the colleagues at the lab for useful feedback; Ferran Marqués, Verónica Vilaplana and Marc Combalia for useful discussions about image processing. All authors acknowledge the support of the Spanish Ministry of Science, Innovation and Universities to the EMBL partnership, the Centro de Excelencia Severo Ochoa, and the CERCA Programme / Generalitat de Catalunya.

Author Contributions

Conceptualization: Manuel Muñoz-Aguirre, Vasilis F. Ntasis, Roderic Guigó.

Formal analysis: Manuel Muñoz-Aguirre, Vasilis F. Ntasis.

Methodology: Manuel Muñoz-Aguirre, Vasilis F. Ntasis.

Software: Manuel Muñoz-Aguirre, Vasilis F. Ntasis.

Supervision: Manuel Muñoz-Aguirre, Roderic Guigó.

Validation: Manuel Muñoz-Aguirre, Vasilis F. Ntasis, Santiago Rojas.

Writing – original draft: Manuel Muñoz-Aguirre, Vasilis F. Ntasis.

Writing – review & editing: Manuel Muñoz-Aguirre, Vasilis F. Ntasis, Santiago Rojas, Roderic Guigó.

References

1. Abels E, Pantanowitz L, Aeffner F, Zarella MD, van der Laak J, Bui MM, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol. 2019; 249: 286–294. https://doi.org/10.1002/path.5331 PMID: 31355445
2. Parwani AV. Next generation diagnostic pathology: use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. Diagn Pathol. 2019; 14: 138. https://doi.org/10.1186/s13000-019-0921-2 PMID: 31881972
3. Mulrane L, Rexhepaj E, Penney S, Callanan JJ, Gallagher WM. Automated image analysis in histopathology: a valuable tool in medical diagnostics. Expert Rev Mol Diagn. 2008; 8: 707–725. https://doi.org/10.1586/14737159.8.6.707 PMID: 18999923
4. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol. 2019; 20: e253–e261. https://doi.org/10.1016/S1470-2045(19)30154-8 PMID: 31044723
5. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017; 318: 2199–2210. https://doi.org/10.1001/jama.2017.14585 PMID: 29234806
6. Nagpal K, Foote D, Liu Y, Chen P-HC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. npj Digital Med. 2019; 2: 48. https://doi.org/10.1038/s41746-019-0112-2 PMID: 31304394
7. Qaiser T, Tsang Y-W, Taniyama D, Sakamoto N, Nakane K, Epstein D, et al. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. Med Image Anal. 2019; 55: 1–14. https://doi.org/10.1016/j.media.2019.03.014 PMID: 30991188

8. Hou L, Nguyen V, Kanevsky AB, Samaras D, Kurc TM, Zhao T, et al. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. Pattern Recognit. 2019; 86: 188–200. https://doi.org/10.1016/j.patcog.2018.09.007 PMID: 30631215
9. Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat Mach Intell. 2019; 1: 236–245. https://doi.org/10.1038/s42256-019-0052-1
10. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018; 24: 1559–1567. https://doi.org/10.1038/s41591-018-0177-5 PMID: 30224757
11. Banna GL, Olivier T, Rundo F, Malapelle U, Fraggetta F, Libra M, et al. The promise of digital biopsy for the prediction of tumor molecular features and clinical outcomes associated with immunotherapy. Front Med (Lausanne). 2019; 6: 172. https://doi.org/10.3389/fmed.2019.00172 PMID: 31417906
12. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med Image Anal. 2016; 33: 170–175. https://doi.org/10.1016/j.media.2016.06.037 PMID: 27423409
13. Serag A, Ion-Margineanu A, Qureshi H, McMillan R, Saint Martin M-J, Diamond J, et al. Translational AI and deep learning in diagnostic pathology. Front Med (Lausanne). 2019; 6: 185. https://doi.org/10.3389/fmed.2019.00185 PMID: 31632973
14. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: A survey. arXiv. 2019;
15. Gertych A, Swiderska-Chadaj Z, Ma Z, Ing N, Markiewicz T, Cierniak S, et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. Sci Rep. 2019; 9: 1483. https://doi.org/10.1038/s41598-018-37638-9 PMID: 30728398
16. Stritt M, Stalder AK, Vezzali E. Orbit Image Analysis: An open-source whole slide image analysis tool. PLoS Comput Biol. 2020; 16: e1007313. https://doi.org/10.1371/journal.pcbi.1007313 PMID: 32023239
17. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open source software for digital pathology image analysis. Sci Rep. 2017; 7: 16878. https://doi.org/10.1038/s41598-017-17204-5 PMID: 29203879
18. Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics. 2016; 32: 1395–1401. https://doi.org/10.1093/bioinformatics/btw013 PMID: 26755625
19. Gupta R, Kurc T, Sharma A, Almeida JS, Saltz J. The emergence of pathomics. Curr Pathobiol Rep. 2019; 7: 73–84. https://doi.org/10.1007/s40139-019-00200-x
20. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 2020; 369: 1318–1330. https://doi.org/10.1126/science.aaz1776 PMID: 32913098
21. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: A vendor-neutral software foundation for digital pathology. J Pathol Inform. 2013; 4: 27. https://doi.org/10.4103/2153-3539.119005 PMID: 24244884
22. Felzenszwalb PF, Huttenlocher DP. Efficient Graph-Based Image Segmentation. Int J Comput Vis. 2004; 59: 167–181. https://doi.org/10.1023/B:VISI.0000022288.19776.77
23. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45: 1113–1120. https://doi.org/10.1038/ng.2764 PMID: 24071849

6.2.2 Tile extraction for spatial transcriptomics

To perform the experiments, we will start from a set of 23,952 WSIs from 36 different tissues that were considered acceptable for further use from the pathology review (see Fig. C-9 for the size distribution of all 25,446 available WSIs). As we first want to validate the feasibility of the idea of linking gene expression variation with image patterns, we will attempt to demonstrate the computational model with downsampled versions of the WSIs by a factor of 4x. The microns per pixel are consistent across all slide scans, therefore the obtained resolutions are comparable when performing a scaling by the same factor.

Since the histological composition of each tissue is different, this means that the tile extraction pipeline needs to be configured to extract the tiles in a way that allows for difference foreground-background tolerances across tissues. For example, in the case of stomach and skin, we might be interested in preserving the edges of the tissue segments since they contain valuable information: in stomach (Fig. 6-4a), the edges can help determine the specific type of tissue that was captured when obtaining the sample which could be either from mucosal origin (*lamina propria mucosae*) or composed by layers of smooth muscle fibers (*lamina muscularis mucosae*). In the case of skin (Fig. 6-4b), the outermost layer of the epidermis (i.e. the edges of the tissue sample) which is known as *stratum corneum* is recognized as playing a fundamental role in the maintenance of healthy skin, and abnormalities in this layer have been associated with disease states [173].

Although in these cases the segmentation is relatively straightforward as the tissue is solid and strongly distinguished from the background, we also encounter tissues where this does not happen, such as in the case of adipose samples (Fig. 6-4c) where it is more difficult to perform the background-foreground discrimination. With PyHIST, we can tune parameters in the segmentation pipeline to account for these different types of cases: through a combination of image smoothing with gaussian kernels and different tolerance values for merging connected components, we are able to segment all the tissue types available in GTEx. In Fig. 6-4d an example of a

segmented testis sample is shown, with selected tiles marked with a blue cross. All the selected tiles are saved individually to disk. The tiles are extracted with a size of 224×224 pixels, since this is the standard input shape for many CNN models that perform image classification.

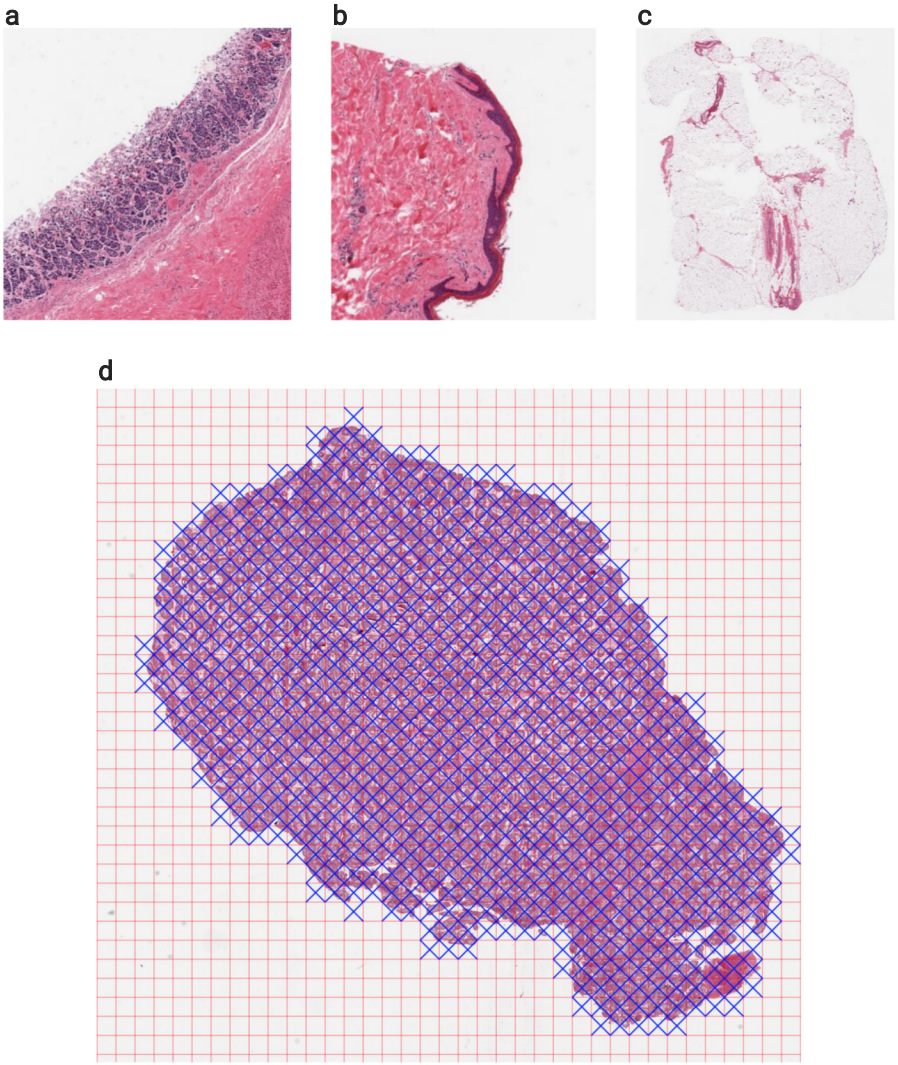


Figure 6-4: **Tile extraction**. **(a)** Subregion of a stomach sample, with the outermost layer corresponding to mucous membrane. **(b)** Subregion of a skin sample, with the stratum corneum layer aligned towards the center. **(c)** Full view of an adipose tissue sample. Panels a-c are shown at different scales only for visualization purposes. **(d)** Segmentation output of a testis sample.

6.3 Feature learning

Once the tiles have been extracted, the next step is compressing them into a lower-dimensional representation in order to be able to link them to molecular traits. Historically, the task of extracting features that describe an image was performed through methods such as template matching, Fourier descriptors, Gabor features, among others [174]. As discussed in Section 2.3, deep learning advances, specifically those based on convolutional operations, allow learning complex characteristics of the input images, and are now generally favored over classical approaches for feature extraction, at least for tasks that require the usage of these features for other downstream goals, such as image classification.

6.3.1 Conceptualizations

The compressed representation can be obtained through several means. One of them is through the use of convolutional autoencoders, which is an unsupervised technique with a neural network consisting of two main parts: an encoder, which transforms the input image into a lower-dimensional representation, and a decoder, which maps this representation back into the shape of the input image, with the loss function quantifying the difference between the input image and its reconstruction. The main difference between a traditional autoencoder and a convolutional autoencoder is that the latter takes into account the spatial structure of the image when computing the feature maps, while a standard autoencoder does not.

Another way to generate vectors of features is by employing pretrained CNNs as fixed feature extractors. Implementations of many network architectures for classification problems exist in deep learning frameworks such as PyTorch [175] or Tensorflow [176], with pretrained weights generated by training each of these models over large collections of images of general, non-domain-specific objects such as the Imagenet database [177]. For a given architecture, when performing the forward pass for an image, we retrieve the generated activations, usually at the last fully connected layer of the network, discarding the softmax and classification part. Note that when using

a model as a fixed feature extractor, the weights of the model are frozen, meaning that no training is performed and the existing (pretrained) weights of the model are used to perform inference. The size of the extracted feature vector will depend on the model architecture, as well as the position in the network from which we decide to retrieve these features.

Lastly, the features can also be extracted through transfer learning and finetuning an existing network. We expose the underlying concept briefly using the formalization introduced in [178]: consider a data domain $\mathcal{D}$ with feature space $\mathcal{X}$ and marginal probability density function $P(X)$ such that $\mathcal{D} = \{\mathcal{X}, P(X)\}$, with $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. A task $\mathcal{T}$ on this domain could be to learn a predictive function f that assigns an instance label $y_i \in \mathcal{Y}$ on the basis of the feature vector x_i, so that $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$. With these definitions, we consider a source domain $\mathcal{D}_s = \{(x_{S_1}, y_{S_1}), \ldots, (x_{S_n}, y_{S_n})\}$, $x_{S_i} \in \mathcal{X}_S$, $y_{S_i} \in \mathcal{Y}_S$ of tuples of data instances and their labels. We define the target domain in the same way, as: $\mathcal{D}_T = \{(x_{T_1}, y_{T_1}), \ldots, (x_{T_n}, y_{T_n})\}$, $x_{T_i} \in \mathcal{X}_T$, $y_{T_i} \in \mathcal{Y}_T$, with the corresponding source and target tasks being $\mathcal{T}_S$ and $\mathcal{T}_T$. Finally, transfer learning consists in improving f_T by reusing information from $\mathcal{D}_s$ and $\mathcal{T}_s$, despite that $\mathcal{D}_s \neq \mathcal{D}_T$ and $\mathcal{T}_S \neq \mathcal{T}_T$.

In our problem, $\mathcal{D}_S$ can refer, for example, to the ImageNet dataset with $\mathcal{Y}_S$ having a large number of classes, while in our problem, the target domain $\mathcal{D}_T$ is specific to histological image tiles, with $\mathcal{Y}_T$ being the label set of human tissues. In deep neural networks, transfer learning is achieved by initializing a model with weights pretrained on D_S and it is trained for some epochs over $\mathcal{D}_T$ (as opposed to feature extraction were no additional training is performed). It has been shown that transfer learning helps in achieving good performances when the set of available data $X_T \in \mathcal{X}_T$ is not large enough to train a complete architecture from scratch [179].

6.3.2 Model training

Here, we will use the last idea to construct the compressed representations for the WSI tiles by framing the feature extraction through a classification problem, as we would like to have these representations to be as different as possible from each other

whenever that makes sense with respect to the shared morphological traits among tissues. To train our classifier models, we sample 50 WSIs from each of the 36 tissues (independently of the WSI having associated RNA-seq data or not) and break them down into tiles using PyHIST, generating a total of 1,075,411 tiles. For each tissue, we partition them into training, validation and testing splits of roughly 70%, 15% and 15%, amounting to 35, 8 and 7 WSIs. Since the classification will be performed at the tile level, all the tiles from a WSI are assigned the WSI-level label. Note that this is not optimal for the task of classification at the tile level, since some tissues might have shared histological components (such as for example, mammary tissue and adipose tissue), and other frameworks exist such as Multiple Instance Learning which can be used in the context of classification to generate a WSI-level label by using bags of tiles and weighting them according to their associations with the class. However, here we are mostly interested in the model as means for feature extraction rather than obtaining the most accurate classification performance at the tile level.

When training the model, we perform *data augmentation*, which refers to a set of transformations applied over the data both to enhance the size of the dataset as well as to improve model generalization (by reducing the chance of overfitting). Such a set of transformations can encompass operations such as color space modifications, image mixing, random erasing, among others [180]. When the transformation is applied over an image, the resulting augmented version of the image is assigned the label of the original image. For training our models, we use the following sequence of transformations (with the probability of each transform ocurring indicated in parenthesis): horizontal flip ($p = 0.5$) $\rightarrow$ vertical flip ($p = 0.5$) $\rightarrow$ shift/scale/rotate ($p = 0.25$) $\rightarrow$ random brightness/contrast ($p = 0.2$) $\rightarrow$ normalize ($p = 1$). We observe that a tile has relatively low probability ($p = 0.15$) of passing unchanged through the sequence of transformations. The exact sequence of transforms applied to each image will vary across training epochs, decreasing the chance of overfitting (since it is not as likely that the exact same version of the image is seen many times). In Fig. 6-5 examples of input tiles are shown on the columns indicated by the label "Original", while the column right to these shows the augmented version of the tiles.

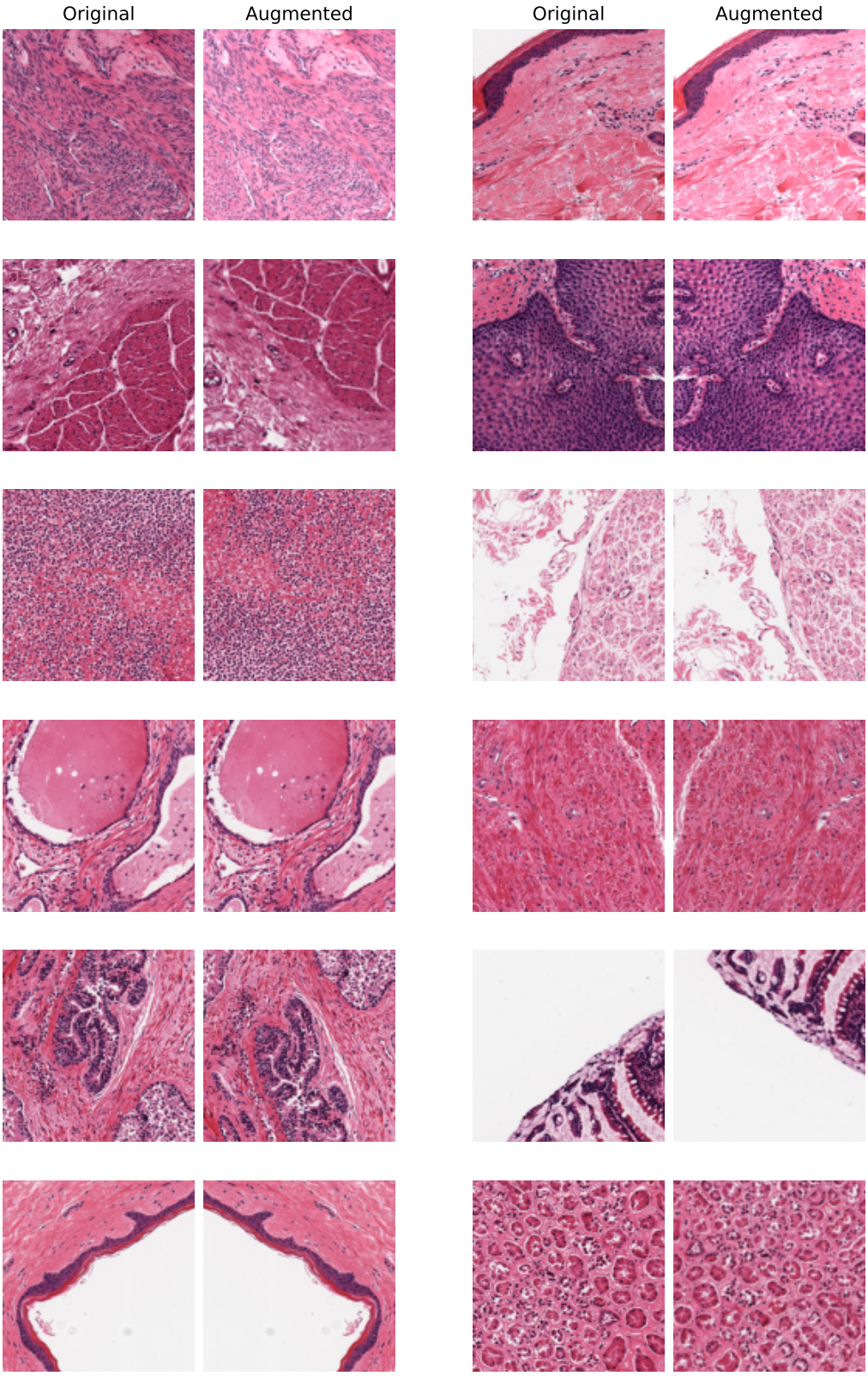


Figure 6-5: **Data augmentation**. Examples of input tiles with their augmented versions, using the sequence of transforms described previously but excluding normalization.

Note that at the end of the chain of transforms, a normalization step is applied: for an input image tensor $\mathbf{X} \in \mathbb{Z}^{N \times H \times W}$ with N channels, we will first elementwise divide it by 255 to normalize it to a $[0, 1]$ range, and then, given a vector of channelwise pixel means $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_N]$ and standard deviations $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_N]$, each channel i is elementwise-normalized as $\mathbf{X}'_{i,\cdot,\cdot} = (\mathbf{X}_{i,\cdot,\cdot} - \mu_i)/\sigma_i$. The specific values used are $\boldsymbol{\mu} = [0.485, 0.456, 0.406]$ and $\boldsymbol{\sigma} = [0.229, 0.224, 0.225]$ since these values correspond to the means and standard deviations of the source task (ImageNet) [181].

Five architectures are independently fit in order to verify that the model performance metrics are robust enough for the same training-validation-testing split: ResNet-18, ResNet-50, ResNet-101 [182], VGG-11 with batch normalization [183], and Densenet-121 [184]. Training is performed in a distributed manner (with the parallelization container `DistributedDataParallel`, see [185] and the PyTorch documentation for details) using the NVIDIA Collective Communication Library (NCCL) across at most 10 GeForce RTX 2080 Ti GPUs. Mini-batch sizes used for the above architectures are 96, 96, 64, 48, and 48, respectively. In a distributed setting, several copies of the training process (at least as many as GPUs that are going to be used) are spawned and initialized under the same conditions, except for the `DistributedDataLoader` object which is specific to each rank (process) and which is used to handle how the tiles are split into mutually exclusive batches of the corresponding size and assigned evenly across GPUs until the data is exhausted. Note that image augmentation is performed on the fly for each batch (as was mentioned earlier, since this increases the chances that different epochs might see different augmented versions of the same input image). An independent copy of the model and the optimizer is held by each GPU which will process the tile batches assigned to it, then the gradients are synchronized (all-reduced average, which means that no explicit parameter broadcasting is necessary) across GPUs while performing the backward pass, so that the model state will be equal in all GPUs before starting the next training epoch.

For each model, the training is performed in two stages: in the first stage, all the pretrained weights of the model backbone are frozen, while the head is replaced by a

linear (fully-connected) layer with of the appropriate size to map the previous layer to an output vector corresponding to the 36 tissue classes. Cross-entropy loss is then used to evaluate the model output with respect to the true labels for each tile: for example, for a mini-batch of size M, the model output tensor is of shape $(M, 36)$ with each row corresponding to the raw scores for each class for a tile; if we define a row vector as $\mathbf{x}$ and the corresponding true class index $c \in [0, 35]$, then the loss for a tile in the mini-batch can be computed as:

$$\ell(\mathbf{x}, c) = -\log\left(\frac{\exp(\mathbf{x}_c)}{\sum_j \exp(\mathbf{x}_j)}\right) \tag{6.3}$$

The loss is computed for all the tiles in the mini-batch and then reduced to a single scalar by averaging, which is used to compute the backward pass. The first stage of training is performed only for a few epochs, since we want to update the gradients for the new fully connected layer put in place. In the second stage of training, we unfreeze all the parameters of the model and resume training normally for a larger number of epochs. Early discussions about layer freezing were presented in [179], which has led to variations in the way that models can be trained when performing transfer learning such as the two-stage process used here, for which some libraries (for example, see [186]) have developed convenience functions. In all of the experiments performed here, optimization is performed with the Adam, a stochastic objective function optimization algorithm based on first-order gradients [187].

At the beginning of each stage, we try to estimate a reasonable starting learning rate for the optimizer by creating a dummy copy of the model and optimizer, defining a lower and upper bound for the learning rate and an exponentially-spaced grid of learning rate trial points between these two boundaries (a variation of the work first described in [188]). A small subset of the data is then evaluated using the learning rates in this grid. For example, in Fig. 6-6, in the range of small learning rates $(10^{-6} - 10^{-3})$ we see that the loss is stable, meaning that it is difficult for the model to perform any learning; from $10^{-3} - 10^{-2}$ the loss starts decreasing quickly, while towards the end of the curve the loss explodes due to the learning rate being too large

which means that the gradient updates might also be very large and thus likely to lead to divergence. The point marked in red shows an example of a good candidate for a starting learning rate (here, chosen as the steepest gradient).

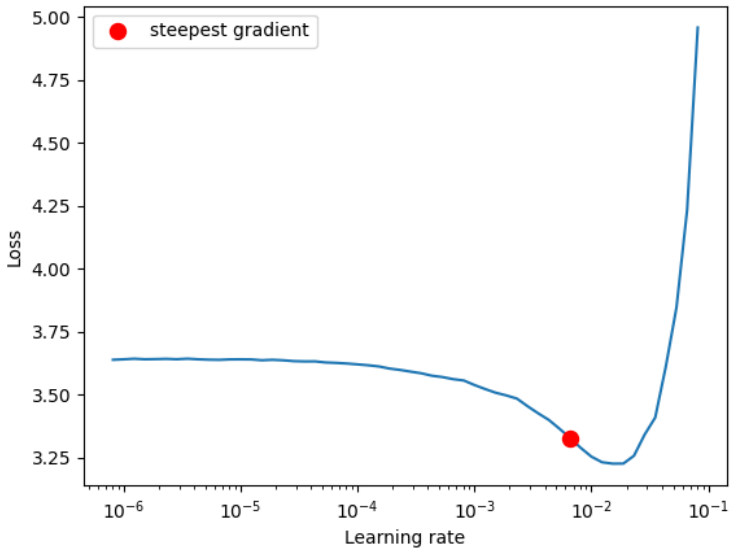


Figure 6-6: **Learning rate tuning**. Learning rate finding for the first training stage of ResNet-50.

In each training epoch, we assess the losses for the training and validation sets (see Fig. C-11a). In general, after 20 epochs, the train loss stops decreasing, with the validation loss following the same trend, but with a gap in the loss between both curves (this will be described later). Model accuracy for the five architectures at the tile level over the test set, which is completely independent of the training-validation process, is similar (77.41%, 78.01%, 76.81%, 77.69%, 78.51%, respectively for the architectures mentioned previously), indicating that there is a stable zone of accuracy which is limited by the nature of the data. From now on, we focus on describing the results of ResNet-50, since it provides a good tradeoff between model complexity (in terms of number of parameters) and performance.

In Fig. 6-7, we show the row-normalized multi-class confusion matrix, with squares in the diagonal being the tissue classification accuracies and off-diagonal squares being misclassified pairs, with numerical values shown only for combinations > 0.1. Tissues

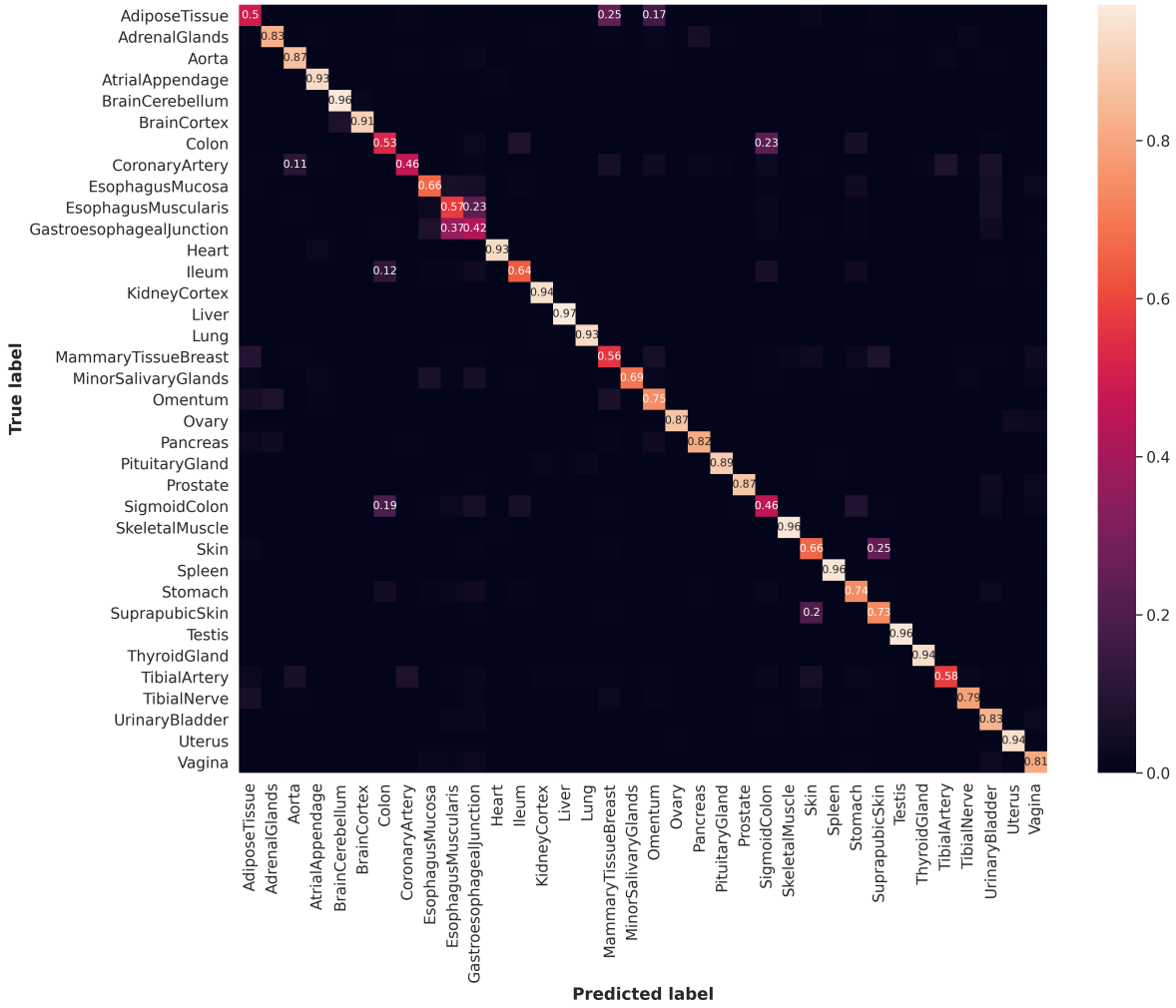


Figure 6-7: **Multi-class confusion matrix at the tile level.** Row-normalized confusion matrix derived by assessing the tile predictions in the test set.

that are homogeneus and unique with respect to their histological features tend to be classified correctly, for example, testis, thyroid gland, brain cerebellum, etc. (see Fig. C-11b for precision, recall and f1-scores), while tissues that fail are those that have shared histological traits, for example, adipose tissue and mammary tissue.

The classification failure for these tissues is due to the fact that the slide label is assigned to all tiles (and thus the term "true label" should be interpreted with care), while these tissues tend to have at least a bivariate composition: for example, mammary tissue also contains a significant amount of adipocytes. The same is true for tissues from the digestive tract, that, despite having specific properties in each tissue, their walls tend to share a basic structure composed by four tunics: mucosa,

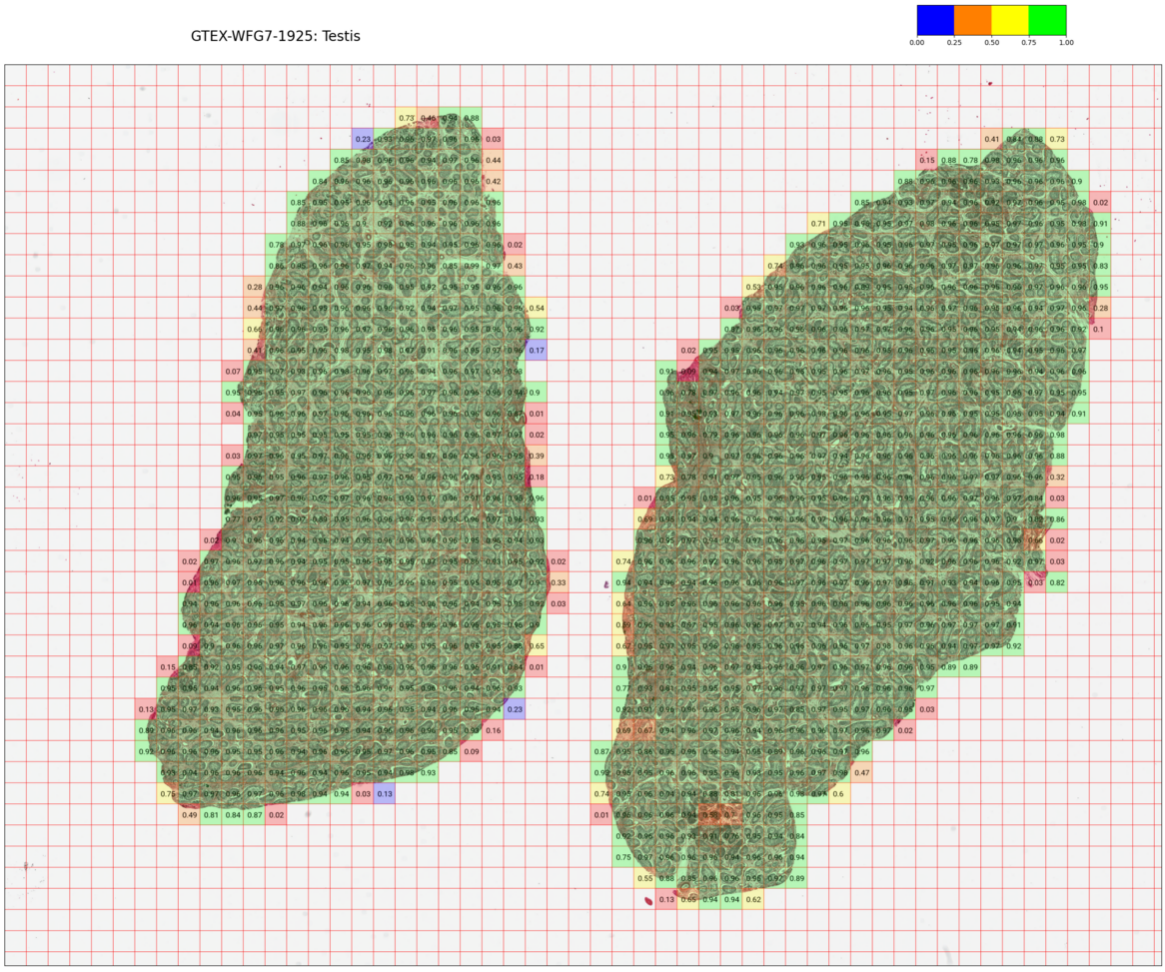


Figure 6-8: **Label-probability visualization for a testis WSI.** The probability with which the true label (testis) was predicted is shown numerically and with a discretized color scale. Tiles in red were predicted incorrectly, while the rest of the tiles are predicted correctly. Squares without a value were considered as background by PyHIST and thus not evaluated by the model.

submucosa, muscularis, and serosa [189]. To demonstrate this, we visualize the prediction probabilities for the slide label for each tile by overlaying the values over the tiled WSI, like in Fig. 6-8 where a WSI for a testis sample is shown and on the top of Fig. 6-9 where a liver sample is shown. Both of these tissues have an overall high prediction accuracy at the tile level across samples since their histological features also tend to be tissue-specific, and the tile prediction failures tend to be concentrated towards the edges of the tissue, for which some of these tiles tend to have a large amount of background content.

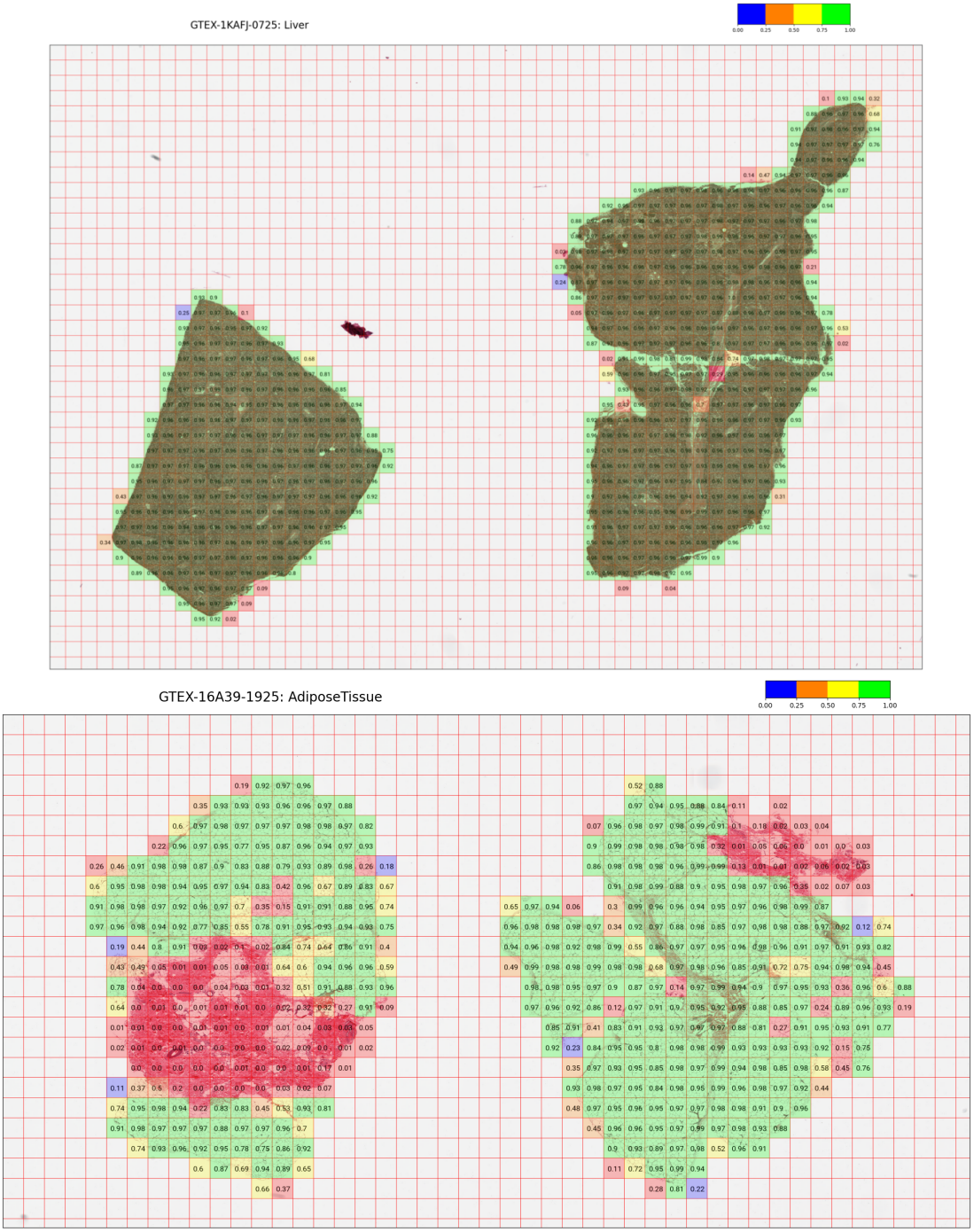


Figure 6-9: **Label-probability visualization for liver and adipose WSIs.** WSIs for liver (top) and adipose tissue (bottom). The probability with which the true label was predicted is shown numerically and with a discretized color scale. Tiles in red were predicted incorrectly, while the rest of the tiles are predicted correctly. Squares without a value were considered as background by PyHIST and thus not evaluated by the model.

Prediction failures, such as some of the tiles in the adipose sample shown at the bottom of Fig. 6-9 tend to occur in heteregenous tissues, as described previously. In the case of this sample, the "incorrectly" classified tiles are connective tissue and not adipose, highlighting the caveat of assigning the WSI-label slide to all tiles. This is also reflected on the gap of the loss functions in training and validation sets described before, as learning cannot be performed successfully for these types of tiles. Nevertheless, here we emphasize that the main interest of building the classifier model is to use it as a feature generator rather than producing the most accurate model. Label predictions at the slide level are verified by assigning the most frequently predicted class across the tiles of each WSI. At the WSI level, accuracy is higher (see Fig. C-12 and C-13 for metrics at the slide level), with most tissues having an accuracy close to 100%, observing that the tissues that tend to fail are also those which fail at the tile level, as expected.

To verify that the activations obtained from the last fully connected layer of the model are as specific to each tissue as possible, we evaluate all the test set tiles (forward pass) and stack their activation vectors into a matrix which is then used to perform dimensionality reduction at tile level. In Fig. 6-10, we observe that tissue identity is preserved with respect to these activations, with tissues with shared histological traits being in the same neighborhood: for example, the two skin types (exposed to sun and not exposed to sun, on the two shades of blue with the inverted 'U' shape) are neighboring, with an interconnecting segment with the types of tiles that tend to be shared between both tissue types. Tissues from the digestive tract as well as those with adipose content (in the center, with shades of brown and orange) tend to have more similarities, and thus also have a less defined identity when compared with tissues that are morphologically very distinct, such as brain (in yellow) or testis (in grey).

In order to verify that the distribution of the learned features is relatively consistent across tiles, for each WSI, we average all the learned features across all tiles, reducing the WSI representation to a single vector of features. We then perform clustering over the correlation matrix obtained from all sample comparisons based

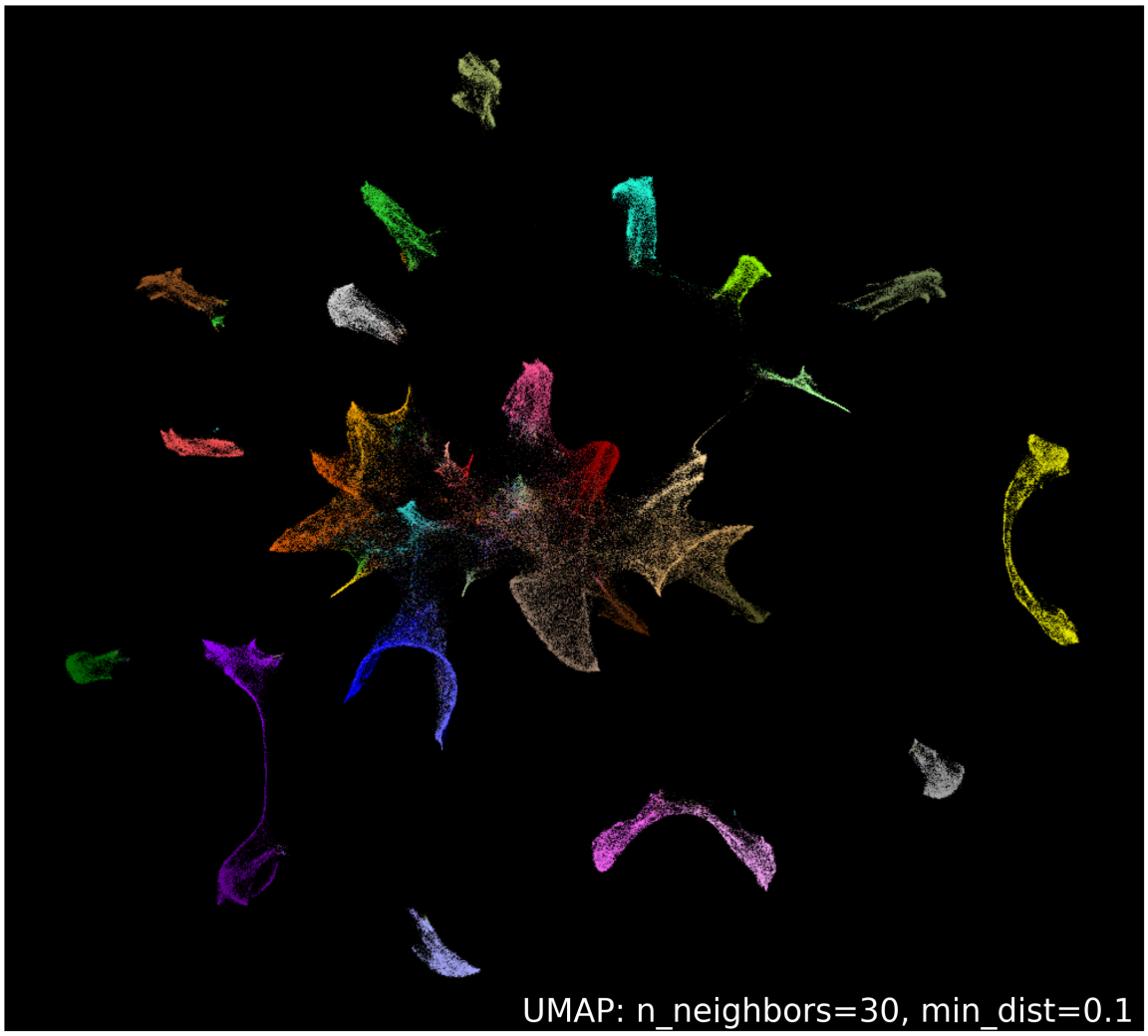


Figure 6-10: **UMAP of tile activations in the test set**. UMAP generated over the stacked matrix of the output vectors from the fully connected layer in the model, for all the tiles in the test set. Each dot represents a tile, color coded by the GTEx tissue convention (see Table B.1).

on their averaged features in order to verify that the WSIs are grouped by tissue identity, as shown in Fig. 6-11. Tissues that had a high prediction accuracy have high sample correlations (in the diagonal), while tissues that have shared features have a correlation substructure as shown in the block in the lower right corner of the figure, exhibiting higher correlations also with samples of other tissues. We now have the feature vectors needed to start performing the association analyses with gene expression. Possible improvements on model training are discussed in Chapter 7.

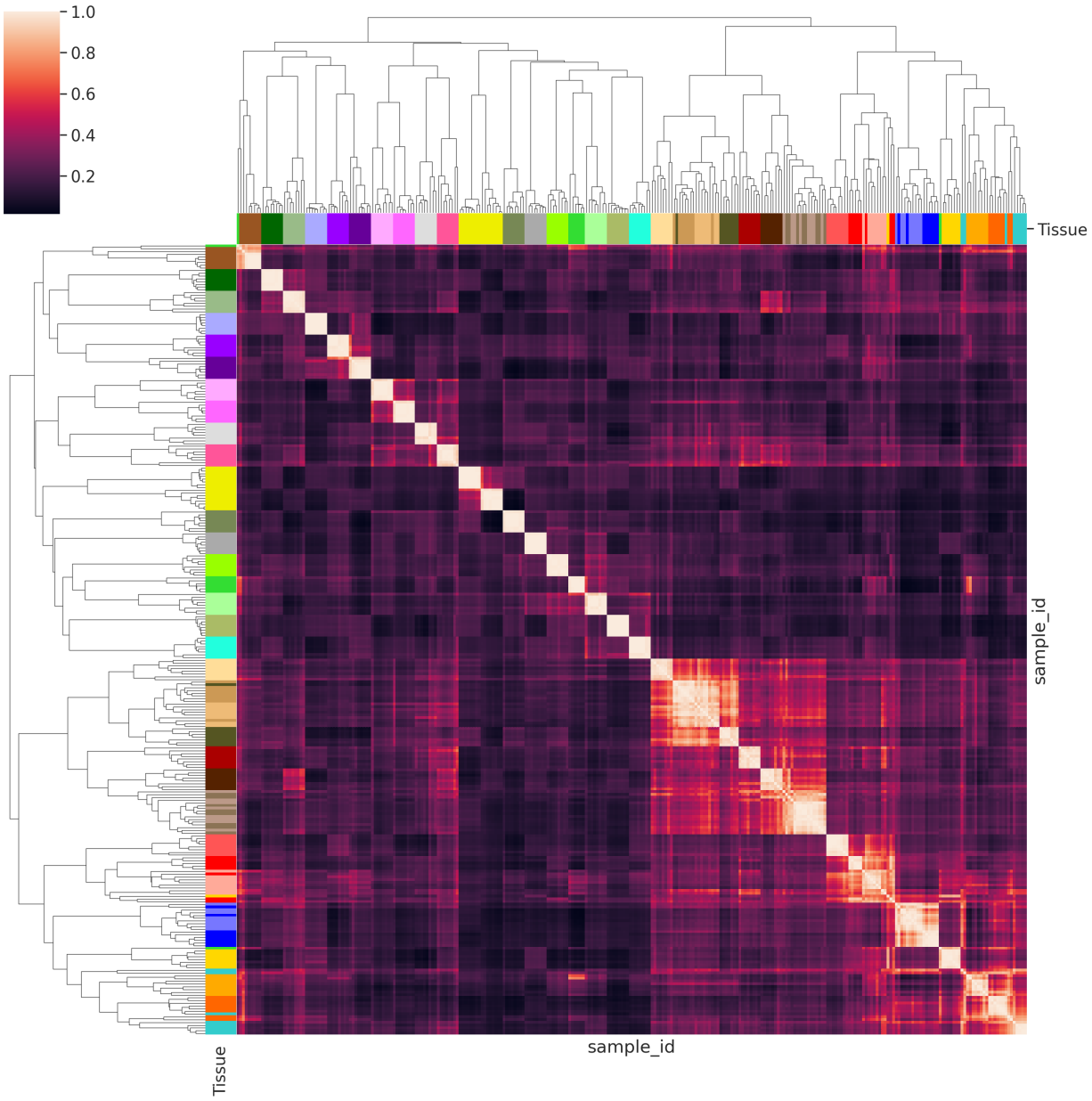


Figure 6-11: **Correlation heatmap of mean activations**. Pearson's correlation is computed between all pairs of samples (across the averaged features at the WSI level), and this matrix is then hierarchically clustered.

6.4 Linking image features with molecular traits

In this section, we briefly describe the current progress on performing the association of gene expression with the learned features. Existing work in the literature has mostly focused on examining multimodal relationships on the basis of correlation, as discussed earlier in Section 2.3.2, and in the case of gene expression, pinpointing specific tiles whose histological features might be related to the expression of that specific gene. Here, we instead aim to spatially deconvolute bulk RNA-seq gene expression and visualize it over the complete WSI. We begin by exploring the idea of how to attribute features in the predicted feature vectors to specific parts of the input image, with the assumption that tiles are summarized with an aggregating function at the WSI level, in such a way that there are one-to-one relationships between WSI feature vectors and bulk RNA-Seq gene expression vectors for a given sample. The need to do this arises because if we identify (gene, image feature) association pairs, we would like to see which parts of the image are associated with that gene, and as such, it is necessary to disentangle what the specific image feature is identifying in the input image.

Although it is not straightforward to interpret a neural network model when compared to classical stastical and machine learning models, algorithms based either on gradients and perturbations have been recently developed to aid in model prediction debugging and interpretability (see for example, [190] which is a compendium of interpretability methods, and used to perform the experiments described here). In general, these methods seek to attribute the output of the model to the input features: if we consider a classifier model $\phi : \mathbf{x} \rightarrow [0, 1]$ with an input $\mathbf{x} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^n$, an attribution is then a comparison of the input vector $\mathbf{x}$ with respect to a baseline input $\mathbf{x}'$, $A_\phi(\mathbf{x}, \mathbf{x}') = [a_1, a_2, \ldots, a_n]$, with each component a_i definining the contribution of x_i to $\phi(\mathbf{x})$. The baseline input should be set so that the prediction is neutral and can be domain specific, for example, a tensor filled with zeros, in the case of an image. Here, we focus on the specific case of attributions computed through integrated gradients [191], which refers to the cumulation of the gradients along all the points in the

straightline path from $\mathbf{x}'$ to $\mathbf{x}$. The integrated gradient for the i-th component in the input, and with $\frac{\partial \phi(\mathbf{x})}{\partial x_i}$ being the gradient of the model output along that component, is defined as:

$$\mathrm{IG}_i(\mathbf{x}) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial \phi(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} d\alpha \tag{6.4}$$

with α being a scaling coefficient. In practice, this can be approximated through a Riemman sum with m steps:

$$\mathrm{IG}_i(\mathbf{x}) \approx \frac{(x_i - x_i')}{m} \sum_{k=1}^{m} \frac{\partial \phi(\mathbf{x}' + \frac{k}{m}(\mathbf{x} - \mathbf{x}'))}{\partial x_i} \tag{6.5}$$

Note that here, integrated gradients are computed for each component in the input with respect to the model output, but we are interested in examining the output of a specific neuron, and thus, it suffices to compute the integral of the gradients with respect to the neuron of interest (which, in our case, should be ideally correlated with the expression of a given gene) instead of the model output.

For a given WSI, we generate the attributions for each tile for one of the neurons with highest layer conductance [192] (which can be interpreted as a way to measure neuron importance with respect to other neurons) in the fully connected layer. The tile attributions are computed through neuron integrated gradients for each channel, obtaining a tensor with the same shape of the input data (i.e. $224 \times 224 \times 3$) for each tile, indicating the relevance of each pixel for that neuron. Then, all the attribution tiles are stitched together and overlaid over the WSI. In this way, for a given neuron, we can visualize what parts of the image are relevant for a given neuron (feature), establishing a basis to spatially resolve gene expression, at least for those (gene, image feature) pairs that are highly correlated. In Fig. 6-12 we show an example of tile attribution computation for all the tiles of the same testis WSI displayed in the previous section, observing that the chosen neuron is mostly capturing the gaps between seminiferous tubules.

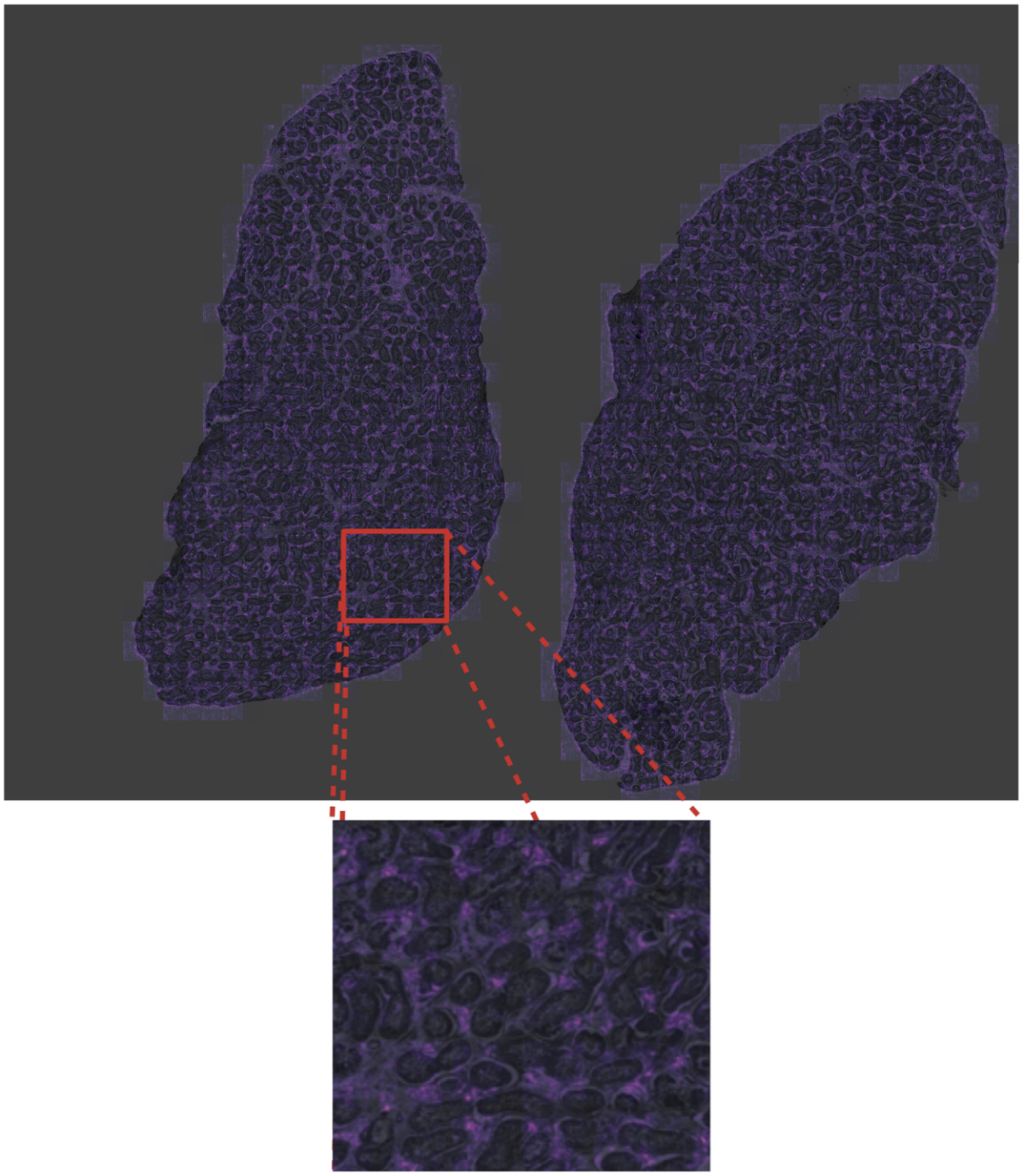


Figure 6-12: **Neuron attributions in a testis sample**. The attributions (shown here in purple) for a specific neuron are computed for all tiles through neuron integrated gradients.

In this attribution example, we see that the image feature captured by the neuron is spatially consistent, despite that the tile attribution evaluations are independent. At the time of writing, we are examining the neuron visualizations in order to determine if they have enough resolution and specificity to serve as a proxy for gene expression visualization. As mentioned earlier, this is an approach based on the assumption that correlation is a good measure to capture associations between image features and expression. As an alternative conceptualization of the spatial deconvolution of gene expression problem, we also propose to treat it under the framework of prediction: it has been shown in the literature that attention mechanisms can be useful to derive the contribution of each tile in classification problems in the context of Multiple Instance Learning (MIL) [193]. In fact, derivatives of the MIL classifier framework incorporating attention mechanisms have already been shown to be of use to determine WSI regions that are useful for class discrimination [194]. We are working towards extending the MIL framework with attention mechanisms to the setting of regression, in order to predict vectors of gene expression from bags of tiles, while examining the specific contribution of each tile to the prediction.

It is worth mentioning that the spatial expression of several sets of genes has already been determined experimentally in the literature, and in our own work (in Chapter 4) we have described sets of genes linked to the muscularis and mucosa layers in stomach tissue. Through these resources, we plan to validate the results of the computational in-silico spatial transcriptomics pipeline.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 7

Discussion and conclusions

7.1 Distilling high-dimensional spaces to examine human phenotypes

With the rapid advances in sequencing technologies from the perspectives of speed, cost and resolution, it has been possible to generate large troves of biological data. Coupled with statistical learning tools, we are able to generate insights about human biology, especially with respect to molecular traits and their relationship with variation in human phenotypes. In this thesis, we examined transcriptome-phenotype relationships with a bottom-up approach through four case studies, linking high-dimensional biological components at different orders of complexity: from RNA, to cell types and finishing with histological texts and images.

In the first case study, presented in Chapter 3, we investigated transcriptional changes in bulk RNA-seq samples as a function of post mortem interval (PMI), and showed that these changes are tissue-specific, with variation in how these are exhibited across time and with most tissues having shifts in gene expression a few hours after death. Our observation has been acknowledged in posterior literature in the context of single cell data, for example, when discussing concerns in sample handling and the possible impact on clustering analyses [195], as well as when characterizing tissue stability after preservation [196] and even when designing guidelines for the logistics

of autopsy programs [197].

We discussed how PMI-related changes could confound gene expression analyses, highlighting the importance of taking into account PMI as a covariate. This consideration can also be extended for studying related data types, for example, RNA-protein correlations [198]. Post-mortem transcriptional changes can have consequences up to the level of assessing drug targets, as discussed in a study where the authors hypothesize about possible effects of variability in gene expression on the effectiveness and safety profiles of drug targets [199].

Based on PMI-dependent gene expression variability, we developed computational models to estimate PMI for a given individual and propose a concept of a protocol that could be used in an hypothetical forensic scenario. Our study focused on demonstrating how tissues could potentially carry the footprint of the time since death. We also hypothesized on whether the cause of death could have an effect on the transcriptome, but acknowledge that we cannot carry out exhaustive analyses on this due to sample size limitations. Subsequent work in the literature has evaluated the effect of the cause of death through spectroscopy in animal models, finding different sensitivities to the cause of death across organs and discussing how these might affect PMI estimation [200], although with a sample size that is still not enough to draw a solid conclusion, and how this generalizes to humans is still unknown.

We emphasize that there are several areas of improvement that need to be addressed in order to fully determine the viability of the PMI prediction protocol for a real application. The cohort we used to study post-mortem changes in expression has a limited PMI timeframe, and thus, we do not have a clear picture of these effects after longer times since death across all tissues, which also means that we cannot make inferences about the reliability of the gene expression-based PMI predictions beyond the range that we studied. Nevertheless, our work has already served as a building block and point of comparison for further literature seeking to improve the reliability of PMI prediction, either through other data types such as metabolomics [201] and composition of microbial communities [202] as well as for studies examining the thanatotranscriptome through samples with longer PMI in specific tissues [203].

In the second case study, presented in Chapter 4, we analyzed primary cells from the ENCODE project together with GTEx data to describe how cells in the human body cluster together in five major cell types based on their transcriptional profiles: endothelial, epithelial, mesenchymal, blood and neural. These types have a close but not exact correspondence to the histological types traditionally defined in the literature to classify organs and tissues. We also analyzed single cell RNA-seq gene expression from 20 mouse organs and tissues and found that most cells and cell types cluster into these five major cell types, providing a degree of evidence on how these transcriptional programs might generalize to mammalian organisms. Further examination of the transcriptional landscape of other organisms is necessary to confirm this hypothesis, especially with regards on characterizing how often specialized cell types might fall outside the observed major cell types, since specialized cell types are likely to have a particular transcriptional signature, as we observe in our work with melanocytes in human and with hepatocytes in mouse.

Using marker genes corresponding to each major cell type, we estimated cellular enrichments in normal tissue samples (from GTEx). We found that each tissue has a specific cell type enrichment signature, which recapitulates gene expression-based tissue organization. Due to the way the enrichments are calculated, a limitation is that they should be interpreted only as a point of comparison between samples within the same major cell type rather than across cell types, and thus, they cannot be strictly considered as a composition. Nevertheless, we found that the enrichment values strongly correlate with estimates derived from using deconvolution methods, such as constrained least squares or support vector regression-based methods such as CIBERSORT. We relied on orthogonal validation measures to ensure that the estimated enrichments for the five major cell types were biologically meaningful through the examination of the GTEx histological data, both from the image and text perspectives.

First, through the examination of histological images, we validated the differences we observed in cell type composition in stomach and transverse colon tissues, in which the sampling process might capture different regions from either mucosal or smooth

muscle origin. We roughly quantified the tissue composition with respect to these two layers in stomach tissue and trained a support vector machine over stomach image features to discriminate the presence of these layers, and applied the model over colon tissue, observing that samples are separated almost identically when compared to the separation based on major cell type enrichment scores. This highlights the importance of considering cell type composition when analyzing gene expression, as in some situations it might act as an undesirable confounding variable. Second, we parsed the pathology review comments associated to each sample using a combination of part-of-speech tagging, rule creation and fuzzy string search in order to derive a set of histological phenotype keywords, generating a resource for the community to study the intermediate nature of histological phenotypes with respect to molecular and other traits. Through these annotations, we found that deviations in cell type enrichments are associated with pathological states. The annotations we derived from the free-text pathology comments have already been used in the literature to examine the predictive power of a few machine learning models for phenotypic variation based on the integration of histological images and genomics [204]. In that study, the authors also mention how these histological phenotypes are associated with gene expression variability, for example, in the case of atherosclerosis in tibial artery, in agreement with our observations about differences in cell type enrichments with respect the same phenotype and tissue. They also observe predictive power for phenotype status when using histological images as input of the models, which is likely to underlie changes in gene expression, supporting the in-silico spatial transcriptomics hypothesis described in Chapter 6 and discussed in the next section. With this work, we have hopefully provided an example of how data integration across different modalities can help to understand the relationship between the human transcriptome and phenotypes through intermediate examinations, here, referring to cell type composition and histology. With the rapid generation of single cell data resources, such as the Human Cell Atlas [205], we will soon learn more about the taxonomy of not only global transcriptional programs, but also about the hierarchical organization of specialized cell types and how this relates to gene expression spatialization in tissues.

In the third case study, presented in Chapter 5, we generated a compendium of sex-differentiated effects in gene expression and its genetic regulation. We performed a sex-differential gene expression analysis controlling for the effects of known covariates, including surrogate variables which we showed to be correlated to tissue cell-type composition. We derived posterior estimates of the effect sizes using an empirical Bayes meta-analysis approach and found that the effect sizes of sex-differential expression are small, tissue-specific but ubiquitous across tissues. This observation has already been replicated in an independent study which repurposed the GTEx dataset to explore sex-differential expression in COVID-19 related genes [206].

We found that effect sizes were stronger in X-linked genes as expected. In relation to this, we built gradient boosted tree classifier models in order to predict sex from autosomal and X-linked genes. The latter had high sensitivity and specificity for the prediction in all tissues (as expected, since their effect sizes are also larger in the differential expression analysis), while this was not the case for autosomal genes, in concordance to the observation that the tissue-wide sex-differential autosomal expression effects are generally small. We interpreted the models through Shapley values and found that top predictive autosomal genes are involved in sex-differentiated traits. In addition, we performed cell type enrichment estimation with xCell signatures, which are different from the ones used in Chapter 4, and replicated the findings of histological differences between males and females in breast tissue, as well as differences between males that had annotated with gynecomastia in the pathology review commments.

In our study, we find sex-differentially expressed genes represented in several biological and molecular functions related to disease and clinical phenotypes, reinforcing the already-known relevance of considering sex as a biological variable when trying to understand disease mechanisms, especially in cases where these effects might be tissue-specific. For example, a recent study about androgen regulation and bowel function has acknowledged our observations about the effects of biological sex on gene expression [207], and in fact, sex differences in gut-brain axis disorders are well documented in the literature [208]. On the other hand, there are also situations in

which it is still unclear how sex differences observed at the clinical level within a disorder are reflected at the molecular level. For example, another study explored the effects of sex differences in the brain transcriptome in schizophrenia and found that these effects were small, with their sex-specific findings showing high replication rates with our sex-differential expression analyses [209].

Understanding disease signatures, especially with respect to genetic effects, is still a challenging topic due to the lack of statistical power. This becomes even harder when considering interaction terms, and this is also a limitation we acknowledge in our work for the tests performed to discern the effects of sex on genetic regulation, which we found to be weak. Another point of improvement is the incorporation of environmental factors that are correlated with sex, such as smoking [210], and for which we did not have data to examine here. Despite these limitations, we believe that our study can be a useful resource of sex effects in non-diseased tissues.

7.2 A data-driven approach to human histology

Advances in storage and computation capabilities have incentivized the adoption of slide digitization across pathology laboratories around the world, making it easier for clinicians to share data as well as empowering the development of computer-aided procedures [211] as the ones mentioned in Section 6.2.1, ranging from disease classification to quality assurance. Models to analyze whole slides can also be a tool to resolve conflicts in cases of interobserver disagreement with respect to the evaluation a specific pathology, which can arise due to differences in detection methods especially in the case when a categorial evaluation scale is used [212].

On the other hand, through single cell sequencing, it has been possible in the last years to expand our knowledge about the composition and organization of cell types across tissues in human and other organisms, but what we can learn about the spatial context solely from this technology is limited since the information about the cellular spatial context is lost. To fill the gap, in situ spatial transcriptomics techniques have been recently developed and applied to spatially resolve gene expression across a variable number of cells, visualizing it over the histology of the sample in order to draw inferences. But there still exists data from paired bulk RNA-seq samples with histological images that could be potentially used to deconvolute the spatial assignation of gene expression, that, if successful, could prove a valuable and cost-effective tool to broaden our understanding of the molecular organization of tissues. To this end, in Chapter 6, we propose to replicate the spatial transcriptomics experimental protocol in a data-driven way through the use of matched samples of bulk RNA-seq gene expression and whole slide histological images; in other words, the goal is to associate variation in gene expression with variation in histological patterns. We discuss how this might be achieved through a three step procedure that involves i) data preprocessing, ii) feature extraction and iii) linking variation in image features with variation in gene expression, and present the results of the computational experiments performed up to this point.

The first step is covered in Section 6.2.1, where we highlight the lack of a standard-

ized way to extract tiles from WSIs for machine learning applications, and developed a software tool to address this: PyHIST, a Histological Image Segmentation Tool. We demonstrate through an example with slides from the The Cancer Genome Atlas that PyHIST is able to successfully preprocess the data, which is then used to train a convolutional neural network in the context of a classification problem. We recover the activations of the last fully-connected layer in the model in order to show that these feature vectors cluster together with respect to morphological tissue traits within the context of cancer. Next, we described the considerations for applying the tool over GTEx WSIs in Section 6.2.2. We then presented in Section 6.3 the procedure to train convolutional neural networks to classify the tissue of origin of a tile, which are then used as a proxy to extract compressed representations of the image tiles in the GTEx dataset, similarly to what we showed in the TCGA example, but with a more controlled training framework to scale up to a larger number of tiles. We show that classification at the tile level is accurate (as measured by accuracy and f1-score) and robust for tissues with clear histological patterns such as liver, testis and spleen, but underperforms in tissues with shared traits, such as those belonging to the digestive tract. We discuss how this is a limitation within the context of classification since we assign to all tiles the WSI-level label, but also highlight that we only use the classifier as means for feature vector generation. Through dimensionality reduction of tile activations and hierarchical clustering of mean tile activations, we show that the distribution of image features preserve tissue identity and relatedness whenever appropriate. Finally, in Section 6.4 we present the result of an early exploration on how to examine the input image $\leftrightarrow$ feature mapping through integrated gradients, and mention an alternative framework to directly predict gene expression from an aggregated representation of the image tiles.

We decided to spend a significant amount of time in the first two steps of the pipeline, since both preprocessing and the generation of feature vectors are critical for the stability of the third step. There are several points of improvement and additional questions that we aim to cover in the final version of the study with regards to these. For example, although we have already demonstrated that the fitted classifier models

are able to generalize reasonably within the context of the GTEx histological image dataset through the use of an independent data partition, the final models need to be cross-validated in order to guarantee their robustness, especially for the remaining parts of the analysis with respect to gene expression spatialization. Another measure to be taken is the implementation of early stopping when training the models. The current classification experiments indicate that tissues with clear histological features that are mostly non-overlapping with other tissues already have a good classification performance, suggesting that it might be unnecessary to include further data for those tissues, while the rest of the tissues may not benefit from larger sample sizes either due to the limitations of assigning the WSI-label to all tiles, as discussed previously. However, this does not mean that we cannot further test the generalization of the model by evaluating a larger set of unseen WSIs.

This also raises the topic of model generalization to other sets of H&E-stained WSIs: most large WSI databases focus on diseased tissue whereas here we examine tissues extracted from healthy individuals, but these other slides could still serve as a control, for example, to see that the spatialized gene expression is within the boundaries of the tissue. One consideration for this is that there exists color variation across datasets generated by different pathology laboratories: albeit the color appearance of the WSIs is globally quite similar, variations in the stain are not uncommon. Despite this, recent research has shown that, at least in the case of CNNs, omitting stain normalization affects model performance only in a very limited way, but suggesting that color augmentations are actually more useful in terms of model robustness and generalization [213], and this is a step that can certainly be incorporated in our data augmentation transformations.

Here, we have chosen a specific resolution that is coarse enough to validate the outcome of the pipeline with orthogonal measures, but it might be worth comparing gene expression spatialization performance with respect to the chosen resolution to perform the analyses, since at each downsampling factor, the amount of global information captured by each tile is different: at lower resolutions (i.e. global view of the WSI), we are able to capture structural information that is obviated at higher

resolutions and thus likely to lead to better model performance for the tissue classification problem, as has already been attempted in the literature [214], at the expense of being less interesting in the context of gene expression spatialization.

Through the generation of image features we can also explore questions that gravitate towards the topic of understanding tissue complexity and organization: how different is tissue taxonomy when comparing the histological and transcriptomic perspectives? Although a traditional histological classification of tissues exists as discussed on Chapter 4 and redefined in our work through the proxy of gene expression, using the image features derived in Chapter 6 we could possibly quantify the degree of similarity between the transcriptomic and histological data modalities, for instance, through the comparison of the hierarchical clustering dendrogram derived from median tissue transcriptome profiles against the dendrogram built from image representations at the WSI level.

The analysis of multimodal data, such as the gene expression and histological image pairs examined here, has the potential to open different lines of research in computational biology and biostatistics: recent work has modeled spatial arrangement of nuclei in a WSI through graph convolutional neural networks for classification of disease states with high concordance with observations made by pathologists [215]. One can easily see that these WSI-derived graphs, besides being useful to identify differences between tissue classes, could serve to validate models of cellular composition that are computationally inferred through bulk RNA-seq.

7.3 The road to precision medicine

In this thesis, we have provided examples on how big and heterogeneous data in biology can be used in conjunction with statistical learning to derive biological knowledge. For example, in Chapter 4 we showed how histological images and text annotations can be used together with gene expression to identify disease states and their changes at the molecular level, specifically with respect to computational inferences of cellular enrichments. Multimodal data types, such as gene expression, histological images and medical reports have potential to generate knowledge that can translate into clinical practice by personalizing treatment decisions, in other words, *precision medicine*.

Indeed, in Chapter 5 we have identified sex-differentially expressed genes that are involved in different biological functions as well as some of them having relevance for disease and in clinical phenotypes, highlighting the importance of taking sex into account in the context of studying disease. There is an immediate example of this in the very relevant (and unfortunate) situation of COVID-19 at the time of writing: possible mechanistic explanations for the sex biases present in this disease have already been discussed in the literature (see [216]) with the authors pointing out the relevance of the differences between males and females at the level of treatment and care. Our observations about the effects of post-mortem interval on gene expression have also been referred to within the context of the disease with respect to concerns about RNA viability when examining FFPE tissue samples through RT-PCR [217]. Studies like these serve as a basis to consider how, in the future, when the necessary technology scales up and becomes more accessible, gene expression and other molecular traits such as cell type composition (in conjunction with other covariates) might be analyzed in a personalized way, leading to individualized clinical decisions and treatments with respect to the features and products of an individual's genome.

There is an important thought that should not be obviated before closing this thesis work and it is that of the ethical concerns of using AI and related methods in life sciences and healthcare applications. Although the possibilities of examining *big data* with statistical learning methods are exciting, care must be exerted when developing

models with clinical goals. Even though we do not derive clinical models in this work, it should not go unsaid that whenever there is a combination of modelization (through any statistical method) and biological data, there exists potential for misuse and bias, either explicitly or implicitly. To deploy a model in a production setting, thorough testing should be performed to ensure that performance is comparable across different strata of the population, for example, when considering different combinations of sex, age and ethnicity, and it is often the case that models in the literature are developed with constrained cohorts. This also implies that a consistent regulatory framework for production models needs to be developed (see, for example, the European Commision discussions about AI [218]), taking into account a wide array of possible legal issues that may arise. Many ML-based models that are now prevalently used fall under the category of what is known as a "black box", certainly helping in the task of prediction and function estimation to solve scientific questions, but with the inner workings of the model still being relatively obscure. Interpretability methods, such as the ones applied in this thesis work, aim to provide more transparency in this regard, but their power is currently limited especially when considering the ever-growing size of models in terms of parameters: for example, GPT-3, which is a recently published autoregressive language model, has around 175 billion parameters [219]. Although its capabilities to produce human-like language are impressive, the authors acknowledge that data biases are encoded in the model which can also potentially translate to biases in decision-making with respect to population strata, thus highlighting the need to have balanced data cohorts and careful model design, as well as having cross-disciplinary teams to cover the big picture of the problem at hand.

Statistical learning helps us to understand the nature of the data as long as the necessary precautions are taken, since the generated results also provide educated guesses about strengths and weaknesses to account for in future research, because after all, science is an iterative process.

7.4 Conclusions

This thesis work has performed an exploration on how to link transcriptomic changes with human phenotypes at different orders of magnitude, from RNA, to computationally-inferred cell types and histological images, with four specific contributions:

1. Development of models to estimate human post-mortem intervals based on tissue gene expression.

2. A transcriptional-based definition of major cell types in the human body which correspond broadly to the basic histological types of tissue classification, validated through a bridge between gene expression, computationally-inferred major cell type enrichments, histopathology images and pathology review comments in free-text form, as well as through the independent examination of single cell data of mouse organs and tissues.

3. An extensive chacterization of the transcriptional landscape of sex-differential gene expression, finding that effects are small and tissue-specific, but ubiquitous across tissues.

4. An end-to-end conceptualization to associate variation in bulk RNA-seq gene expression with variation in histological image patterns through matched sample pairs. We develop a software tool for whole slide image preprocessing, and demonstrate its usage through a classification example with slides from The Cancer Genome Atlas. We build CNN models to classify image tiles in the GTEx dataset and extract compressed representations to associate image features with gene expression, and perform an early experiment to visualize tile attributions.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

List of contributions

1. **List of publications covered in the chapters of this thesis, in order of presentation ([*] indicates co-first authorship):**

   This thesis is submitted as a compendium of four articles, which are enlisted here:

   - Ferreira, P. G., **Muñoz-Aguirre, M.**, Reverter, F., Sá Godinho, C. P., Sousa, A., Amadoz, A., Sodaei, R., Hidalgo, M. R., Pervouchine, D., Carbonell-Caballero, J., Nurtdinov, R., Breschi, A., Amador, R., Oliveira, P., Çubuk, C., Curado, J., Aguet, F., Oliveira, C., Dopazo, J., . . . , Guigó, R. (2018). The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature Communications*, 9(1).
     https://doi.org/10.1038/s41467-017-02772-x
     JIF (2019): 12.121, JIF (5-year): 13.610
     In collection: Top 50: Life and Biological Sciences (2018)

   - Breschi, A.*, **Muñoz-Aguirre, M.***, Wucher, V.*, Davis, C. A., Garrido-Martín, D., Djebali, S., Gillis, J., Pervouchine, D. D., Vlasova, A., Dobin, A., Zaleski, C., Drenkow, J., Danyko, C., Scavelli, A., Reverter, F., Snyder, M. P., Gingeras, T. R., Guigó, R. (2020). A limited set of transcriptional programs define major cell types. *Genome Research*, 30(7), 1047-1059.
     https://doi.org/10.1101/gr.263186.120

JIF (2019): 11.093, JIF (5-year): 12.010

- Oliva, M.*, **Muñoz-Aguirre, M.***, Kim-Hellmuth, S.*, Wucher, V., Gewirtz, A. D. H., Cotter, D. J., Parsana, P., Kasela, S., Balliu, B., Viñuela, A., Castel, S. E., Mohammadi, P., Aguet, F., Zou, Y., Khramtsova, E. A., Skol, A. D., Garrido-Martín, D., Reverter, F., ..., Guigó R., Stranger B. E. (2020). The impact of sex on gene expression across human tissues. *Science*, 369(6509).
  https://doi.org/10.1126/science.aba3066
  JIF (2019): 41.845, JIF (5-year): 44.372
  Perspective of this work: Wilson, M.A. (2020). Searching for sex differences. *Science*, 369(6509). https://doi.org/10.1126/science.abd8340

- **Muñoz-Aguirre, M.***, Ntasis, V. F.*, Rojas, S. Guigó, R. (2020). PyHIST: A Histological Image Segmentation Tool. *Cold Spring Harbor Laboratory*.
  https://doi.org/10.1101/2020.05.07.082461
  JIF (2019): 4.7, JIF (5-year): 5.26

2. **Other contributions:**

- Kim-Hellmuth, S.*, Aguet, F.*, Oliva, M., **Muñoz-Aguirre, M.**, Kasela, S., Wucher, V., Castel, S. E., Hamel, A. R., Viñuela, A., Roberts, A. L., Mangul, S., Wen, X., Wang, G., Barbeira, A. N., Garrido-Martín, D., Nadel, B. B., Zou, Y., Bonazzola, R., ..., Guigó R., ..., Lappalainen T. (2020). Cell type–specific genetic regulation of gene expression across human tissues. *Science*, 369(6509).
  https://doi.org/10.1126/science.aaz8528

- Research perspective: **Muñoz-Aguirre, M.**, Ferreira, P. G., Guigó R. (2018). Determinación de la hora de la muerte a partir del patrón de expresión de los genes en múltiples tejidos.
  https://genotipia.com/genetica_medica_news/post-mortem-expresion/

3. **Publications as part of the GTEx consortium:**

   - The GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776
   - The GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. Nature 550, 204-213 https://doi.org/10.1038/nature24277

4. **Conferences and other activities:**

   Talks:

   - LifeTime Unconference. July 22-23, 2019. Barcelona, Spain. Talk: Connecting bulk transcriptomic variation with human phenotypes by integrating RNAseq with histopathology images.
   - ISBER's 2nd Biospecimen Research Symposium: Focus on Quality and Standards. February 5-6, 2019. Berlin, Germany. Talk: The effects of death and post-mortem cold ischemia on human tissue transcriptomes.
   - GTEx Project Community Meeting. June 27-28, 2017. Rockville, MD, USA. Talk: Leveraging the GTEx histological data: phenotype extraction.

   Posters:

   - Biology of Genomes. May 8-12, 2018. Cold Spring Harbor, NY, USA. Poster: Characterizing transcriptomic variation across human phenotypes by integrating RNAseq data with histopathology images and annotations.
   - B-Debate: Enhancing the Usage of Human Genomics for the benefit of all. April 20-21, 2017. Barcelona, Spain. Poster: Leveraging the GTEx histological data: phenotype extraction
   - CRG PhD Symposium. 2017. Barcelona, Spain. Poster: Exploring the relationship between the transcriptome and human phenotypes.

- Biostatnet. 2016. Barcelona, Spain. Collaborative poster: Exploring the relationship between the transcriptome and human phenotypes.

Other:

- Barcelona Citython 2019: Rethinking mobility in cities. Winner of the Comprehensive Cities category. Using deep Q-learning to propose a traffic and pedestrian mobility solution. Work presented at the Smart City Expo World Congress.
- Accenture Digital Healthcare Hackaton 2019. Prediction of mortality rate in melanoma patients: Finalist (4th place), developed a gradient-boosting based survival model.
- Jornadas de Cooperación CONACyT-Catalunya. 3MT Contest (3-minute thesis). Winner of the Barcelona Supercomputing Center prize.
- Barcelona Citython 2018: Winner of the CISCO tech prize. Anonymously identifying crowds of people through deep learning. Work presented at the Smart City Expo World Congress.

Detailed author contributions are stated within each article, as well as in bullet-point summaries in each chapter in this thesis work. Signed, the director, Roderic Guigó.

Appendix B

Supplementary tables

Tissue	Abbreviation	Color	Tissue group
Adipose - Subcutaneous	ADPSBQ		Adipose Tissue
Adipose - Visceral (Omentum)	ADPVSC		Adipose Tissue
Adrenal Gland	ADRNLG		Adrenal Gland
Artery - Aorta	ARTAORT		Blood Vessel
Artery - Coronary	ARTCRN		Blood Vessel
Artery - Tibial	ARTTBL		Blood Vessel
Brain - Amygdala	BRNAMY		Brain
Brain - Anterior cingulate cortex (BA24)	BRNACC		Brain
Brain - Caudate (basal ganglia)	BRNCDT		Brain
Brain - Cerebellar Hemisphere	BRNCHB		Brain
Brain - Cerebellum	BRNCHA		Brain
Brain - Cortex	BRNCTXA		Brain
Brain - Frontal Cortex (BA9)	BRNCTXB		Brain
Brain - Hippocampus	BRNHPP		Brain
Brain - Hypothalamus	BRNHPT		Brain
Brain - Nucleus accumbens (basal ganglia)	BRNNCC		Brain
Brain - Putamen (basal ganglia)	BRNPTM		Brain
Brain - Spinal cord (cervical c-1)	BRNSPC		Brain
Brain - Substantia nigra	BRNSNG		Brain
Breast - Mammary Tissue	BREAST		Breast
Cells - Cultured fibroblasts	FIBRBLS		Fibroblasts
Cells - EBV-transformed lymphocytes	LCL		Blood
Colon - Sigmoid	CLNSGM		Colon
Colon - Transverse	CLNTRN		Colon
Esophagus - Gastroesophageal Junction	ESPGEJ		Esophagus
Esophagus - Mucosa	ESPMCS		Esophagus
Esophagus - Muscularis	ESPMSL		Esophagus
Heart - Atrial Appendage	HRTAA		Heart
Heart - Left Ventricle	HRTLV		Heart
Kidney - Cortex	KDNCTX		Kidney
Liver	LIVER		Liver
Lung	LUNG		Lung
Minor Salivary Gland	SLVRYG		Salivary Gland
Muscle - Skeletal	MSCLSK		Muscle
Nerve - Tibial	NERVET		Nerve
Pancreas	PNCREAS		Pancreas
Pituitary	PTTARY		Pituitary
Skin - Not Sun Exposed (Suprapubic)	SKINNS		Skin
Skin - Sun Exposed (Lower leg)	SKINS		Skin
Small Intestine - Terminal Ileum	SNTTRM		Small Intestine
Spleen	SPLEEN		Spleen
Stomach	STMACH		Stomach
Thyroid	THYROID		Thyroid
Whole Blood	WHLBLD		Blood

Table B.1: **GTEx tissues and their abbreviations**. Through this thesis, we use the abbreviations and color legend in this table to identify each tissue.

Appendix C

Supplementary figures

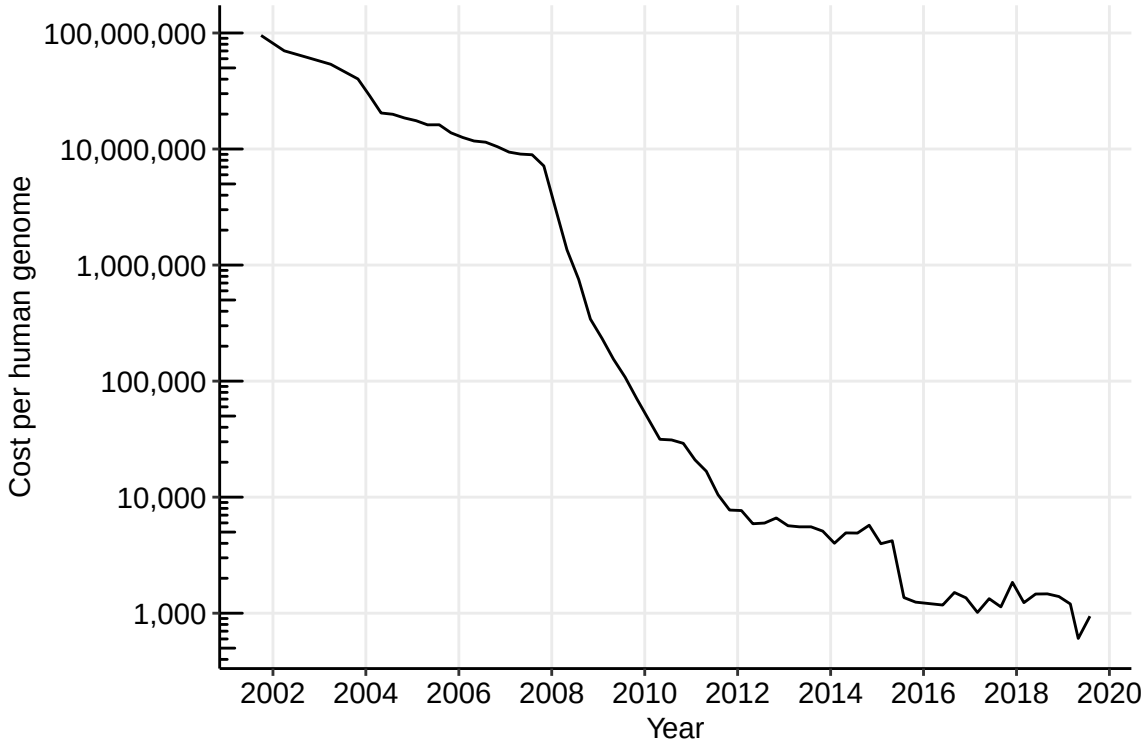


Figure C-1: **Cost of human genome sequencing through time.** The y-axis is the $\log_{10}$-spaced cost of sequencing a complete human genome as estimated by the National Human Genome Research Institute (NHGRI). Adapted from data obtained from [10].

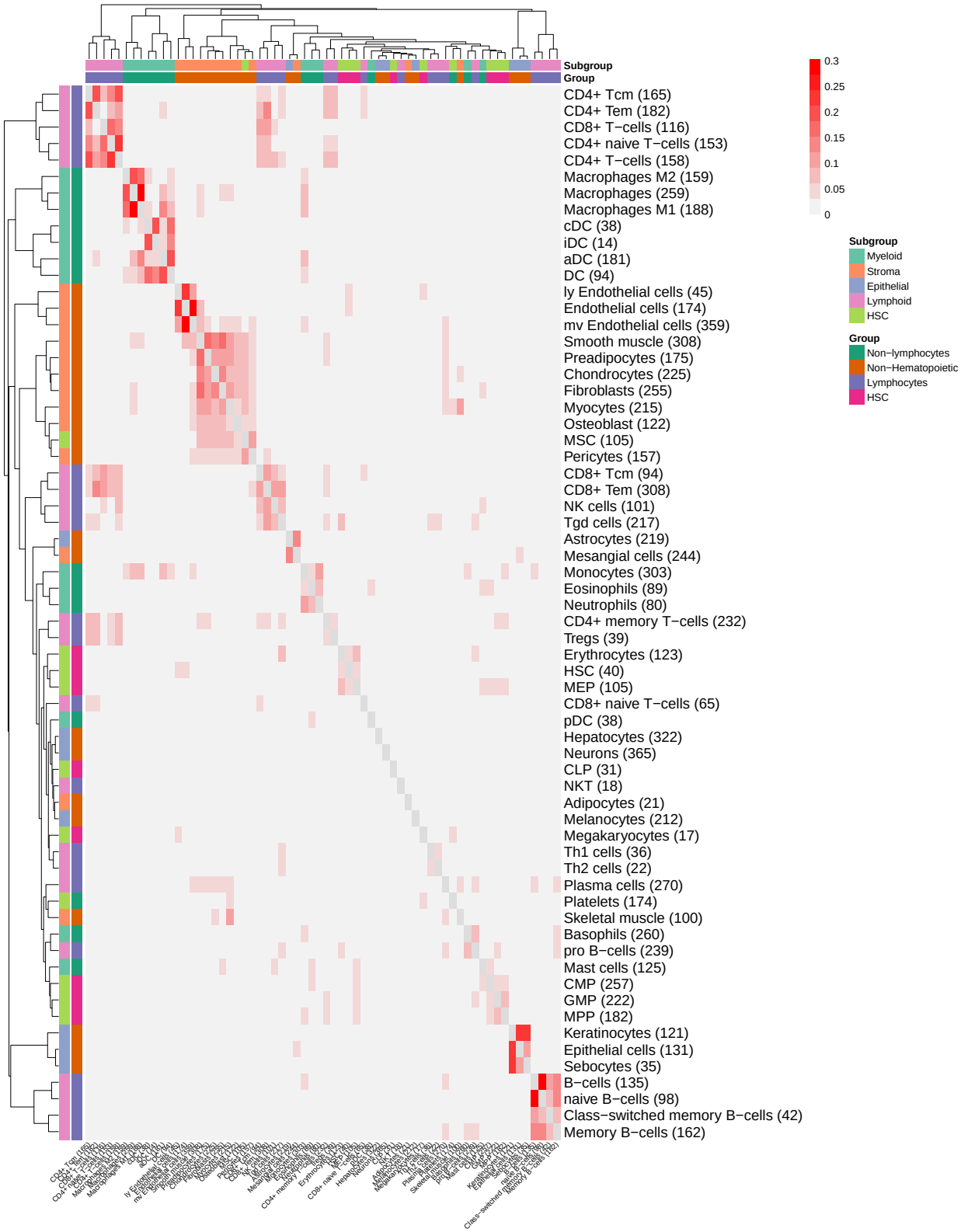


Figure C-2: **Overlap of xCell cell type signatures.** For each cell type, a unique list of marker genes is generated as the union of all marker lists across data sources, with the length indicated in parenthesis. Hierarchical clustering is performed over the matrix of Jaccard indexes for each pair of cell types.

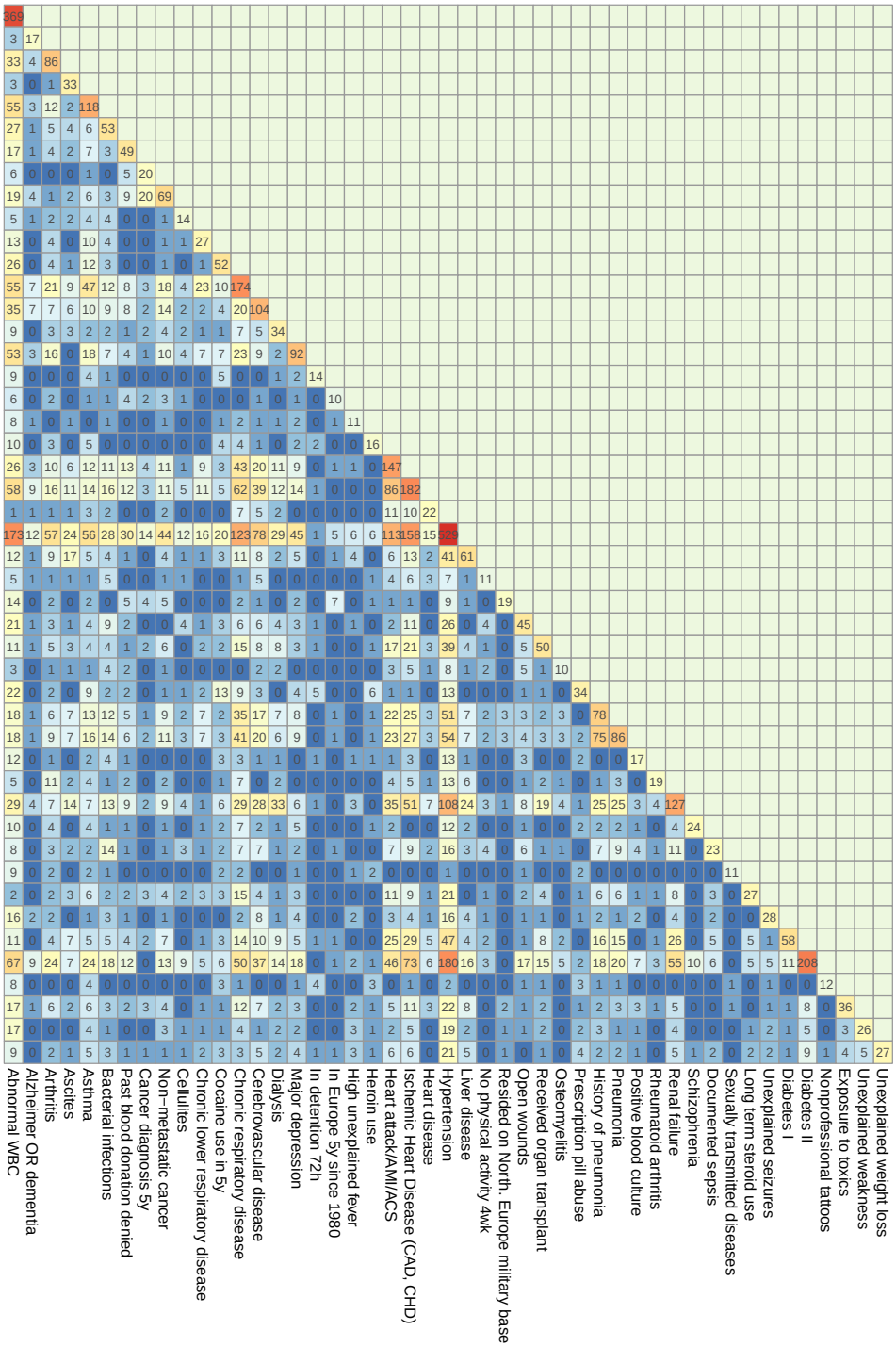


Figure C-3: **Individual medical history.** Co-occurrence of medical history phenotypes across all individuals. Only phenotypes with an incidence larger than 10 individuals were included in the computation of this matrix.

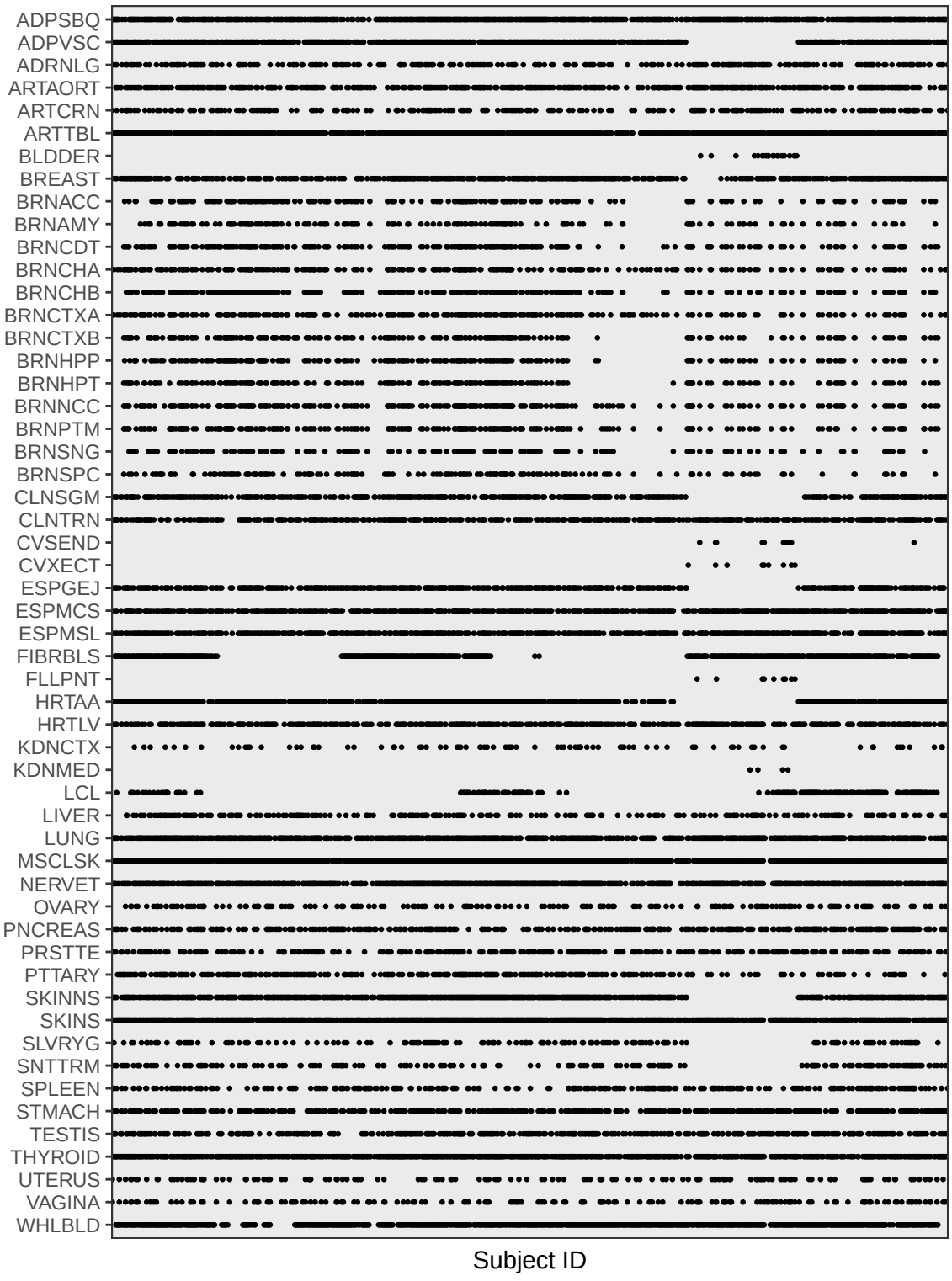


Figure C-4: **GTEx sample availability.** Individual RNAseq sample availability (x-axis) across tissues (y-axis).

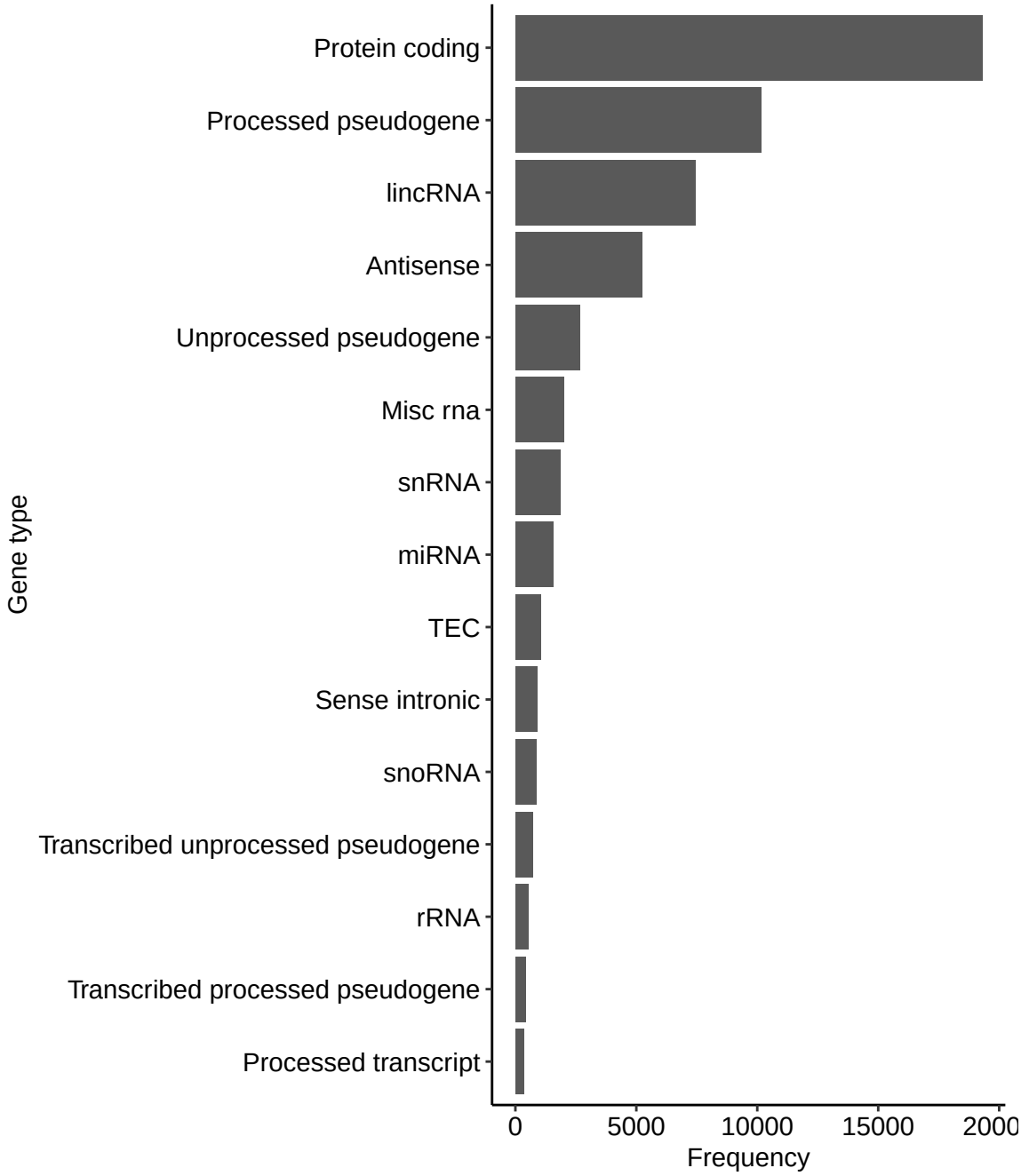


Figure C-5: **GENCODE v26 gene types.** Frequency of the top 15 gene types in the GENCODE v26 annotation.

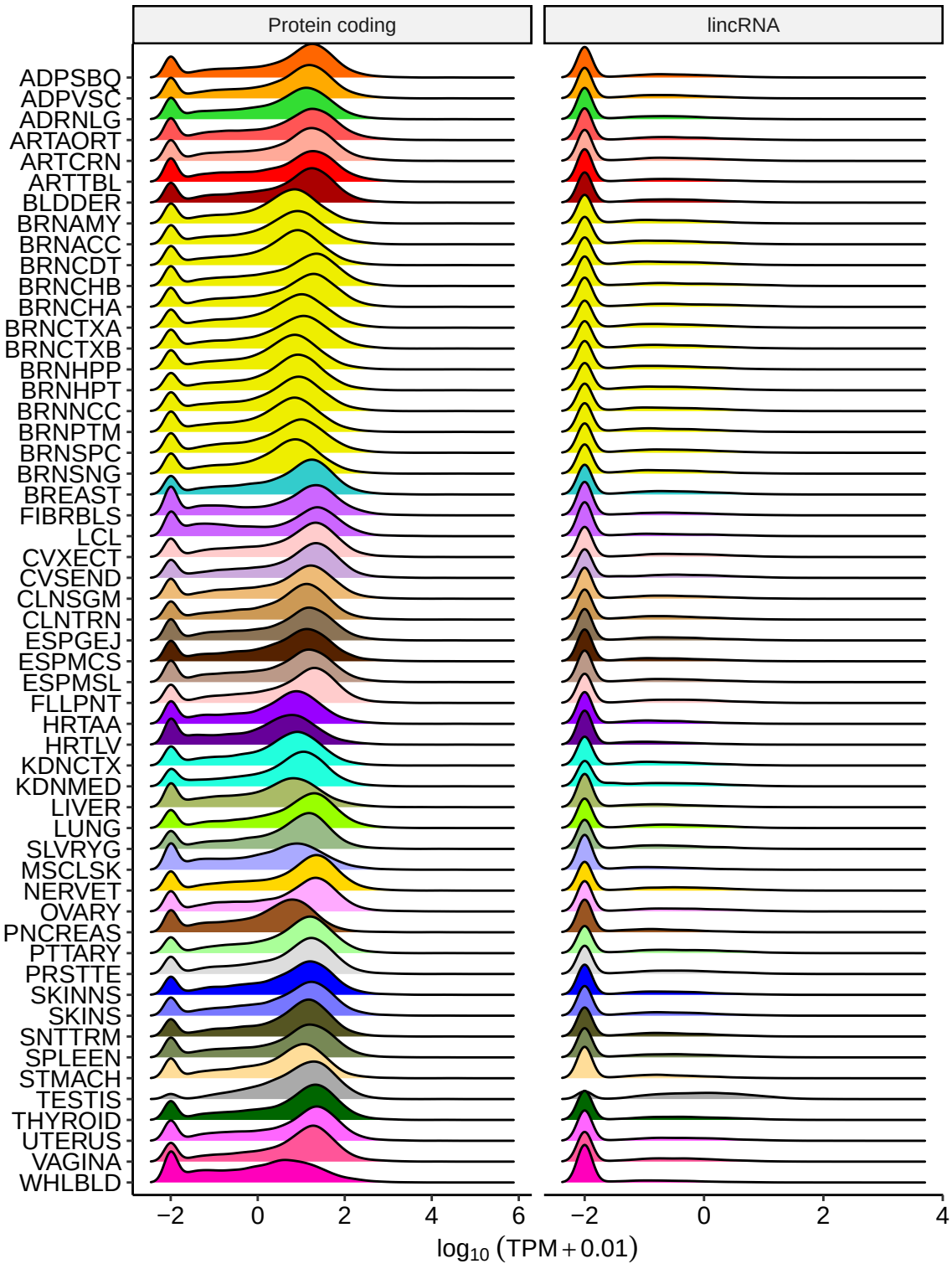


Figure C-6: **Protein coding and lincRNA gene expression distribution.** Density for the $\log_{10}(\text{TPM} + 0.01)$ median gene expression (across samples) per tissue, for protein coding and lincRNA genes.

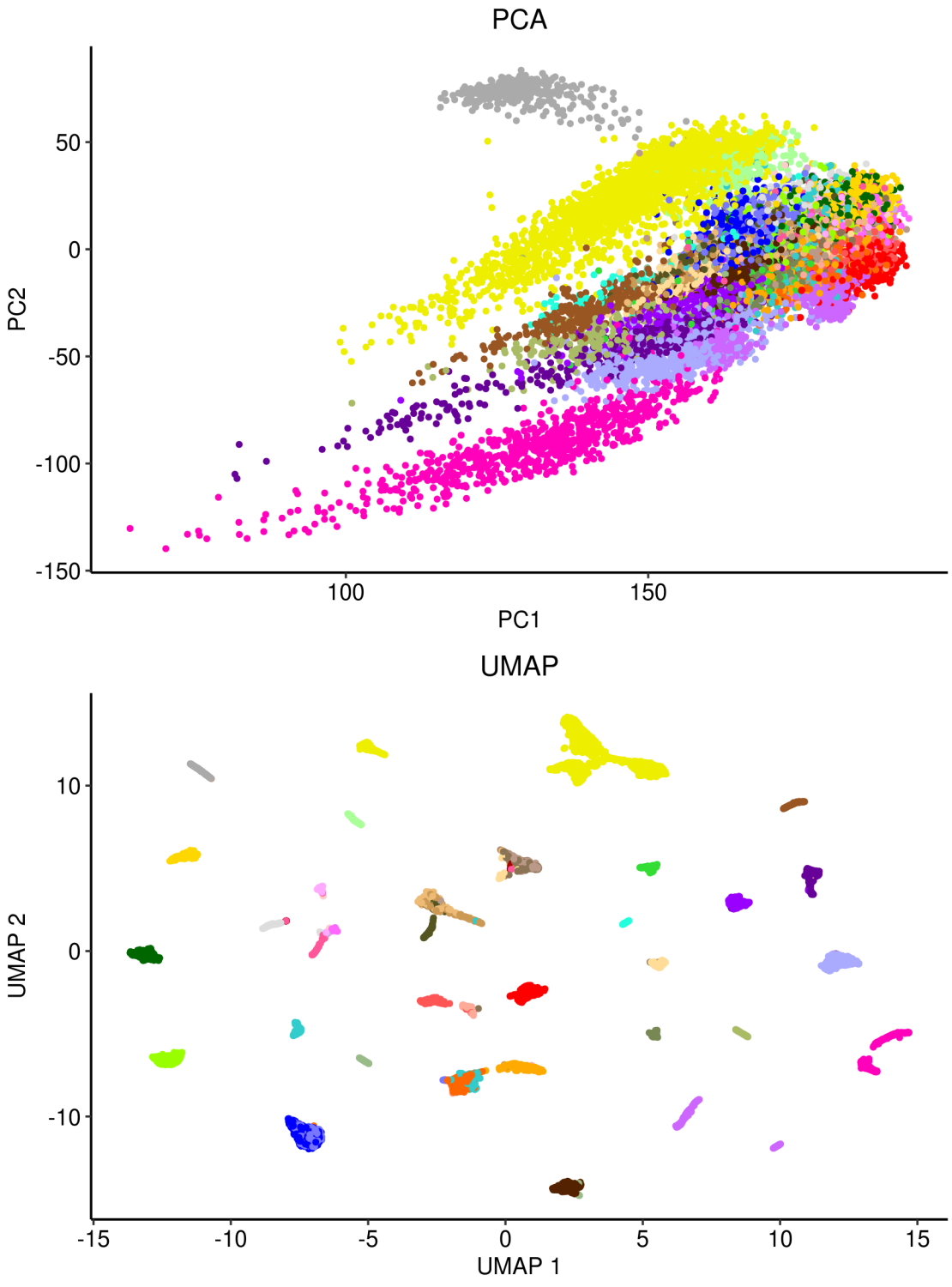


Figure C-7: **Protein coding and lincRNA gene expression distribution.** a) Dimensionality reduction of 17382 RNA-Seq samples based on the expression of protein coding genes, with **(a)** PCA and **(b)** UMAP. Each dot corresponds to one sample.

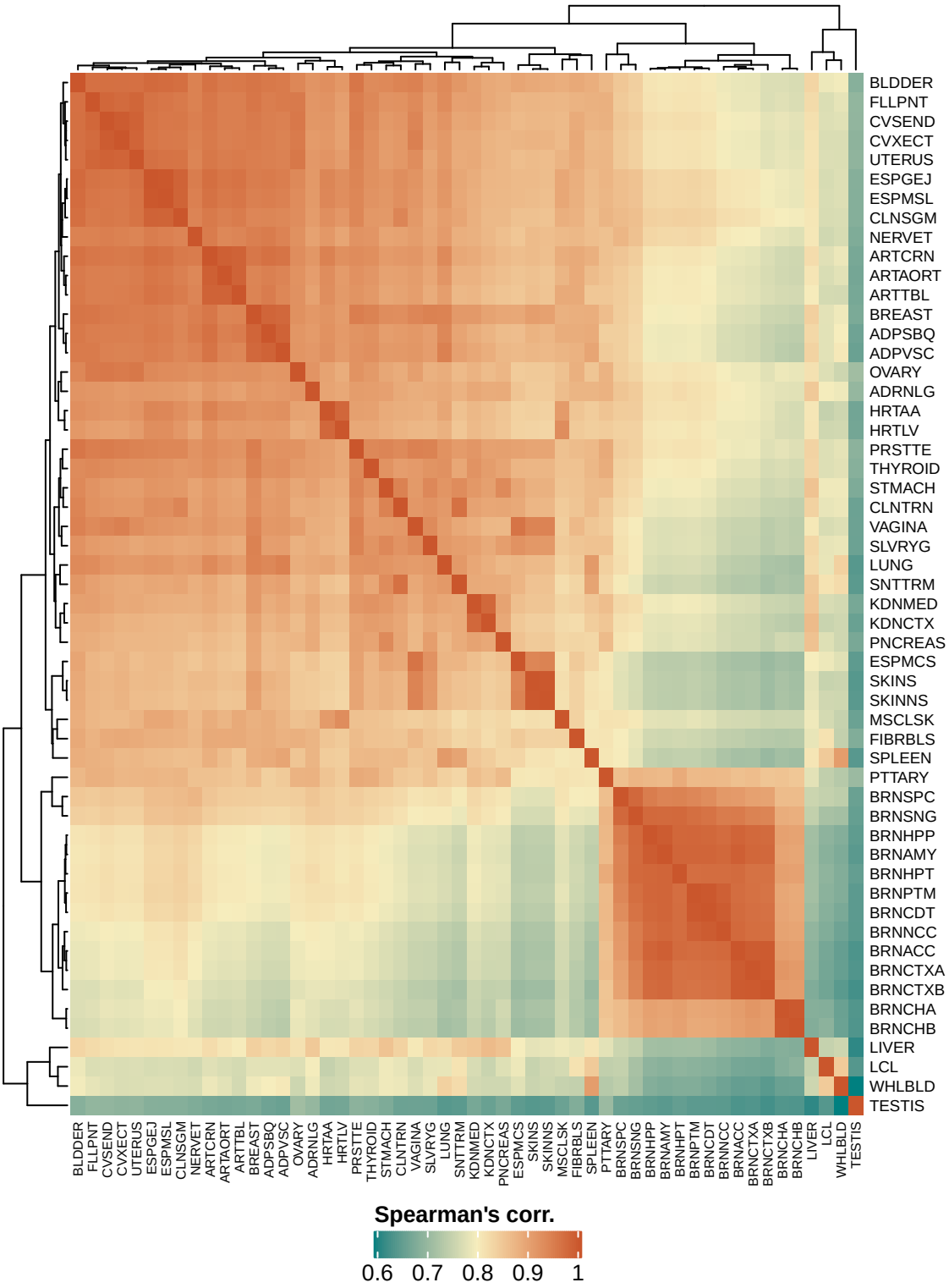


Figure C-8: **Correlation between tissue transcriptional profiles.** Spearman's correlation of median gene expression profiles (across samples) between all tissue pairs. Hierarchical clustering (see section 2.2.1) is performed with complete linkage over a distance matrix calculated with Euclidean distance.

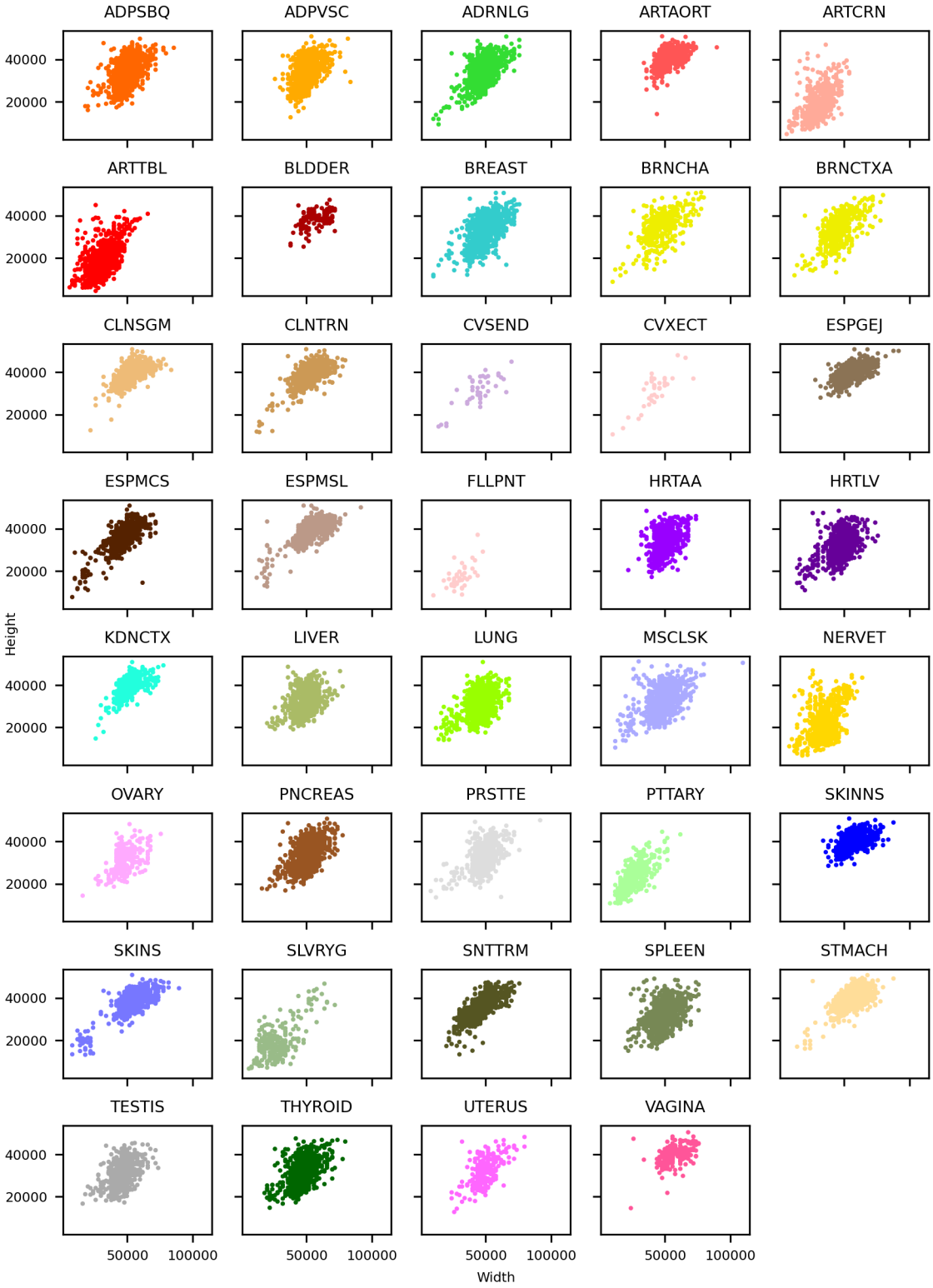


Figure C-9: **Histological image size distribution.** Width × height for the native resolution level of 25,446 histological image slides from 39 different tissues of the GTEx project.

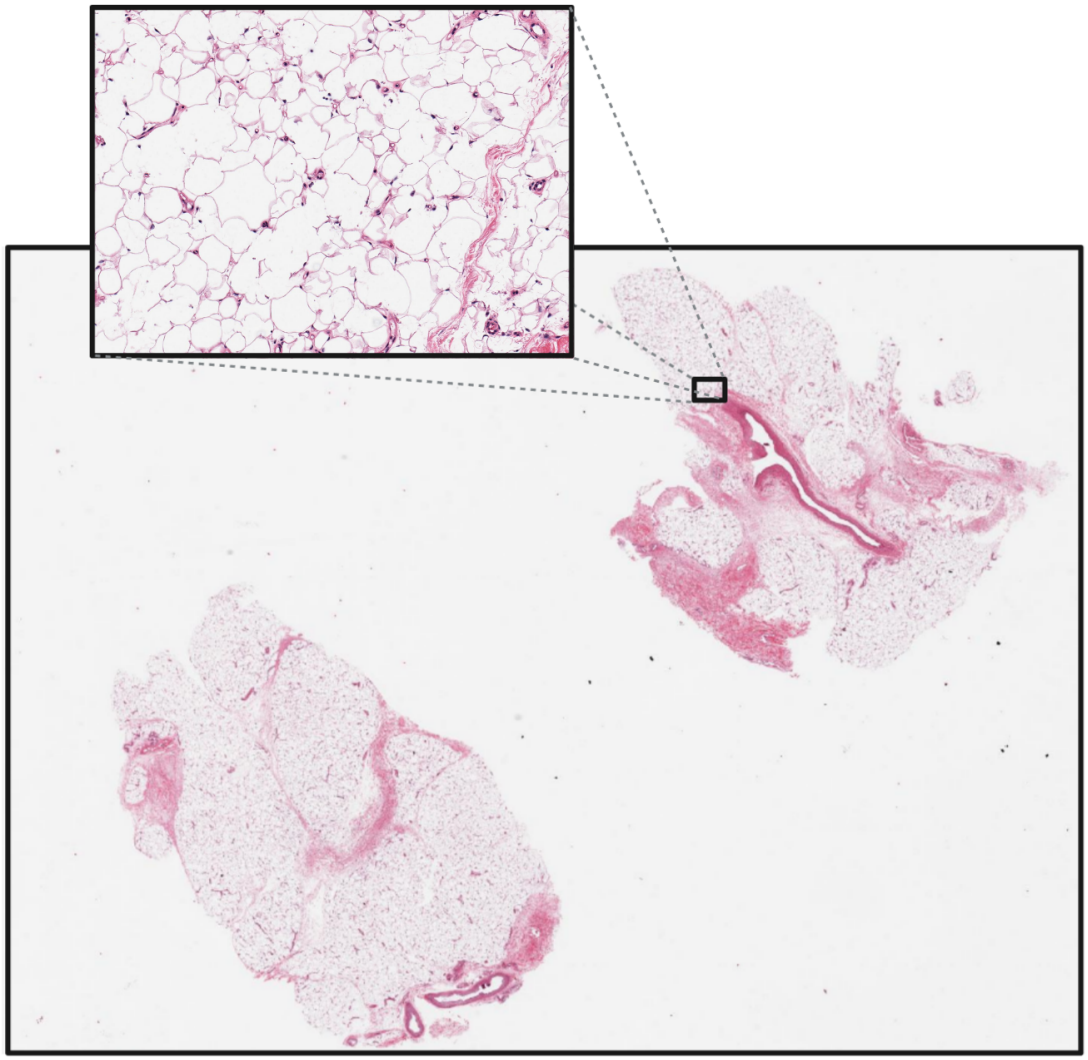
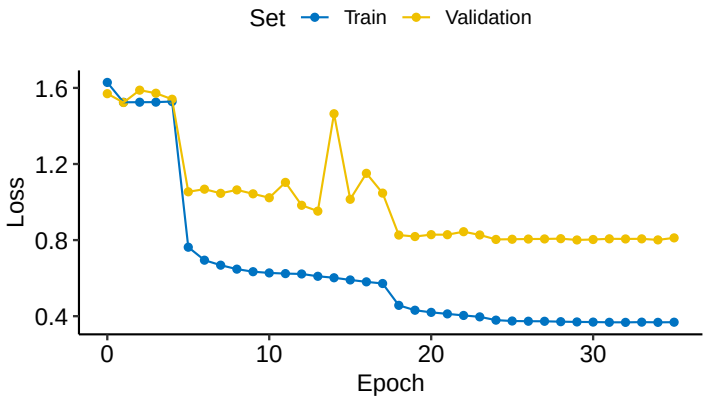


Figure C-10: **WSI snapshot of an adipose tissue sample.** The inset is zooming into a small region of the WSI containing adipocytes, which could potentially be confused with the slide background.

(a)



(b)

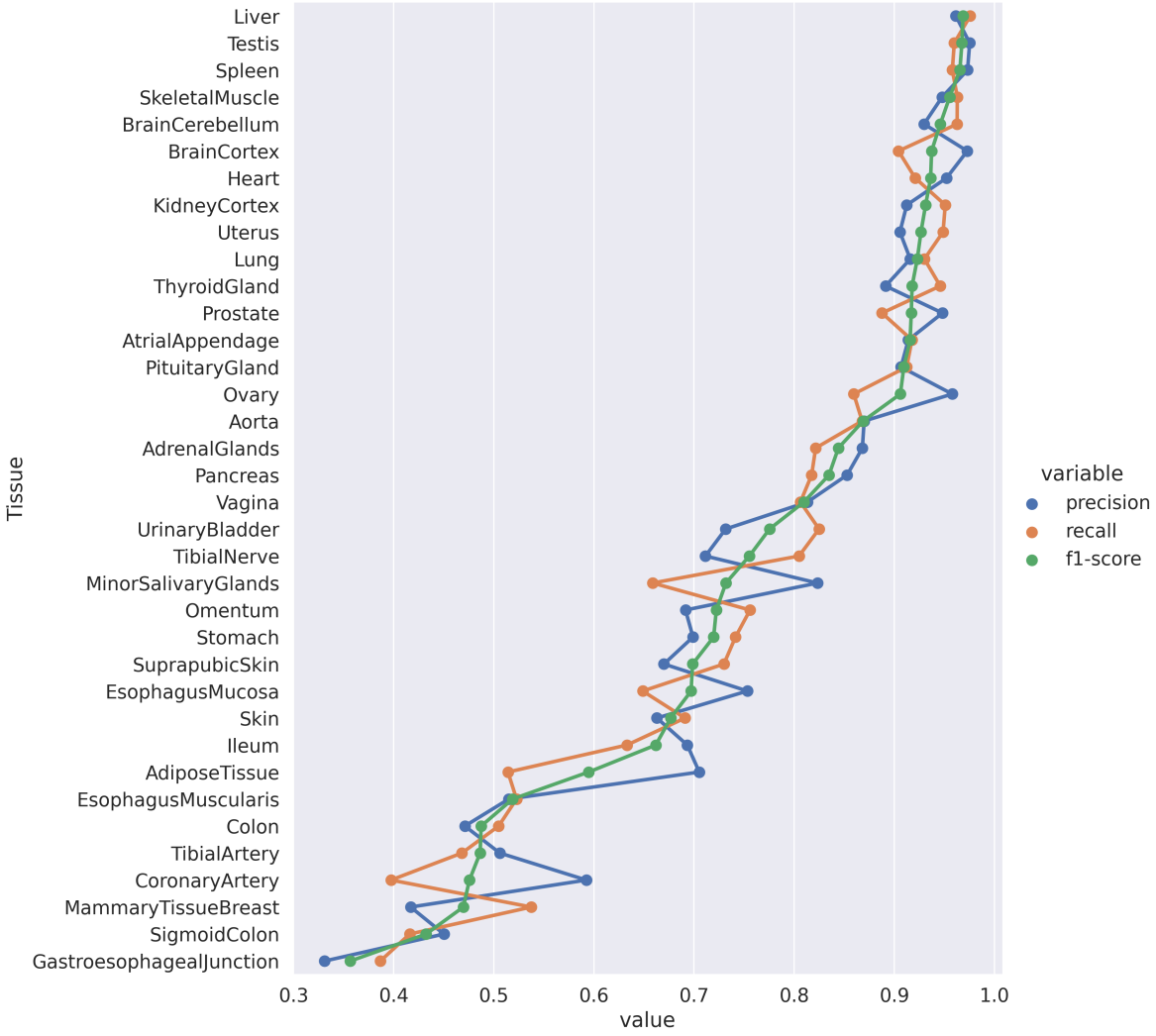


Figure C-11: **ResNet-50 performance metrics at the tile level**. On top, the loss curves for the training and validation sets are shown. The bottom plot shows the precision, recall and f1-scores for the test set tiles in each tissue.

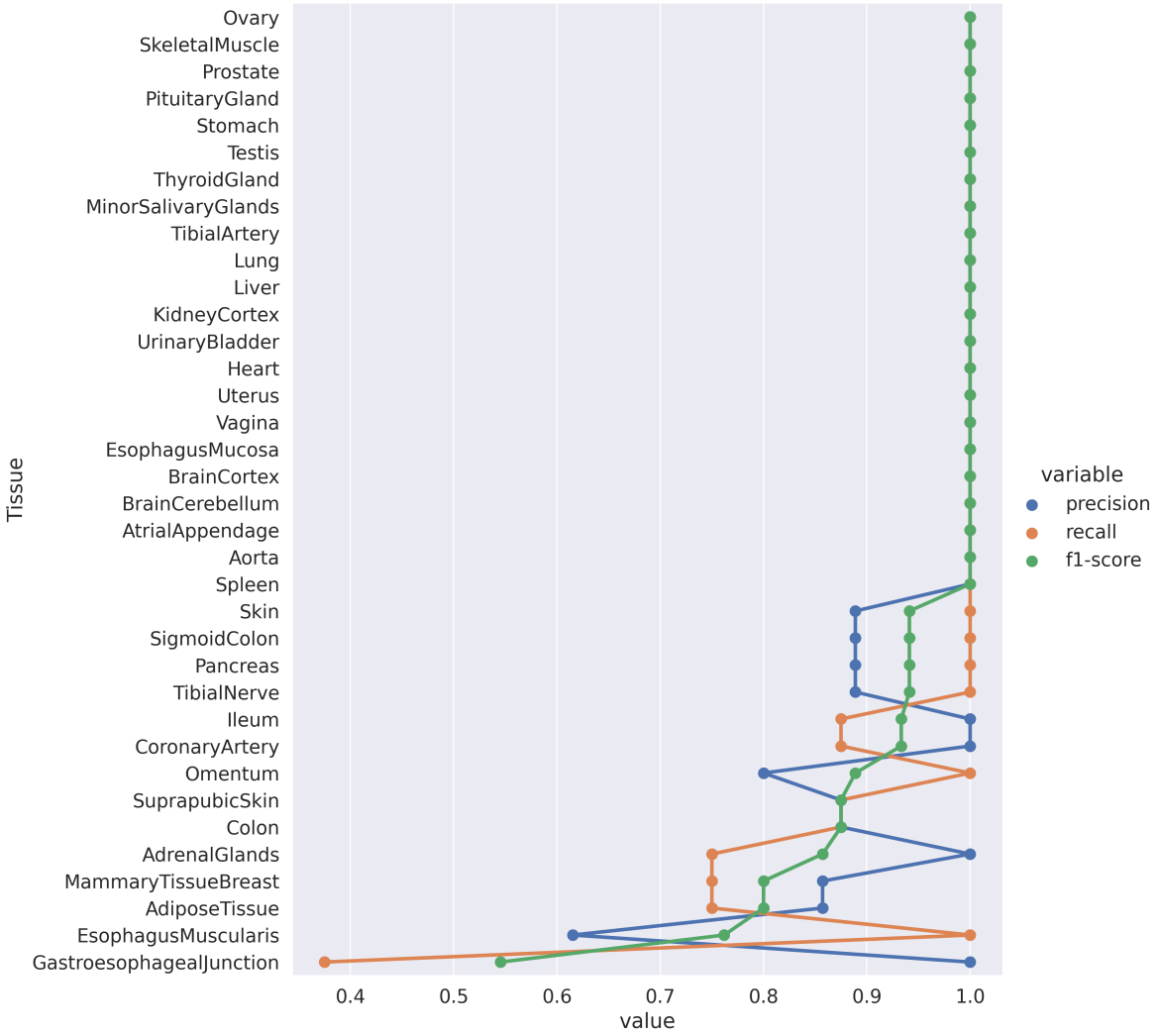


Figure C-12: **ResNet-50 performance metrics at the slide level.** Precision, recall and f1-scores computed at the WSI level by assigning to each slide the label of the most commonly predicted class across test set tiles.

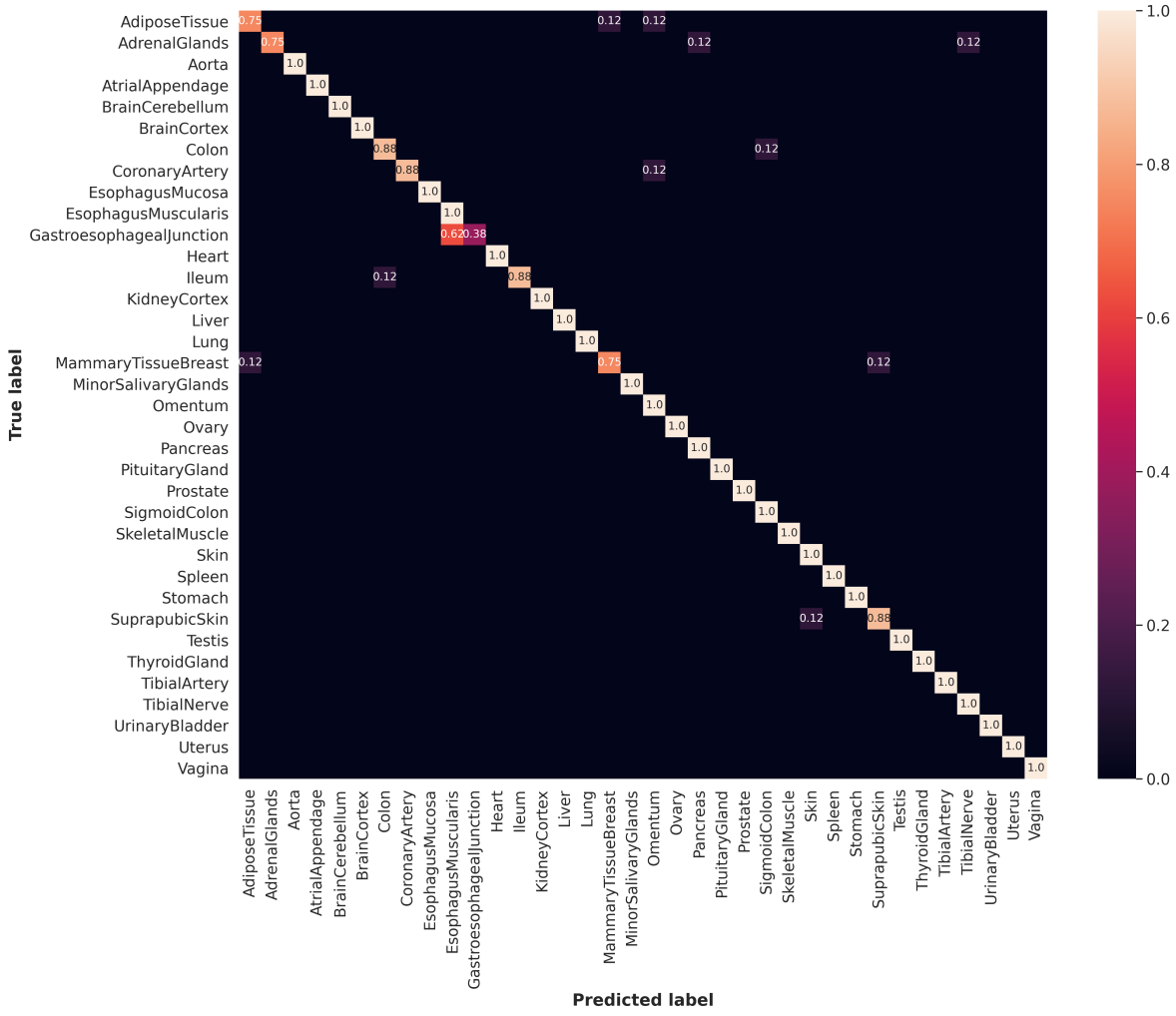


Figure C-13: **Multi-class confusion matrix at the slide level.** Row-normalized confusion matrix derived from the predicted WSI-level labels (by assigning the most commonly predicted class across tiles) in the test set.

Appendix D

Notes

Figures 1-2, 1-11, 1-16, 6-2, 6-3, 6-12 and C-10 were generated (either partially or totally) using BioRender. Figures 2-12 and 2-13 were partially generated with NN-SVG [220]. The rest of the figures were generated with R (ggplot2), Python (matplotlib) and Inkscape. This thesis was written with `emacs` using LaTeX and is based on the MIT thesis template. All the computational analyses carried out in this thesis work were performed in Linux-based distributions, with computing resources provided by the Center for Genomic Regulation (CRG) and the Barcelona Supercomputing Center (BSC). The research carried out in this thesis work was supported by the FPU programme (Formación de Profesorado Universitario) from Ministerio de Educación, Cultura y Deporte (MECD) with predoctoral fellowship FPU15/03635.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

1. Taylor, P. & Lewontin, R. in *The Stanford Encyclopedia of Philosophy* (ed Zalta, E. N.) Summer 2017 (Metaphysics Research Lab, Stanford University, 2017).
2. Frommlet, F., Bogdan, M. & Ramsey, D. in *Computational Biology* 9–30 (Springer London, 2016). https://doi.org/10.1007/978-1-4471-5310-8_2.
3. Orgogozo, V., Morizot, B. & Martin, A. The differential view of genotype-phenotype relationships. *Frontiers in Genetics* **6.** https://doi.org/10.3389/fgene.2015.00179 (May 2015).
4. Hunter, D. J. Gene–environment interactions in human diseases. *Nature Reviews Genetics* **6,** 287–298. https://doi.org/10.1038/nrg1578 (Apr. 2005).
5. Manolio, T. A. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine* **363** (eds Feero, W. G. & Guttmacher, A. E.) 166–176. https://doi.org/10.1056/nejmra0905980 (July 2010).
6. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47,** D1005–D1012. https://doi.org/10.1093/nar/gky1120 (Nov. 2018).
7. Crick, F. *On Protein Synthesis* Accessed: 18-06-2020. http://libgallery.cshl.edu/items/show/52220.
8. Kulski, J. K. in *Next Generation Sequencing - Advances, Applications and Challenges* (InTech, Jan. 2016). https://doi.org/10.5772/61964.
9. Park, S. T. & Kim, J. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. *International Neurourology Journal* **20,** S76–83. https://doi.org/10.5213/inj.1632742.371 (Nov. 2016).
10. Wetterstrand, K. A. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)* Accessed: 18-06-2020. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.
11. De Magalhães, J. P., Finch, C. E. & Janssens, G. Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions. *Ageing Research Reviews* **9,** 315–323. https://doi.org/10.1016/j.arr.2009.10.006 (July 2010).

12. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63. https://doi.org/10.1038/nrg2484 (Jan. 2009).

13. Group, T. S. F. S. W. *Sequence Alignment/Map Format Specification* Accessed: 18-06-2020. 2020. https://samtools.github.io/hts-specs/SAMv1.pdf.

14. Baker, M. De novo genome assembly: what every biologist should know. *Nature Methods* **9,** 333–337. https://doi.org/10.1038/nmeth.1935 (Mar. 2012).

15. Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome Biology* **11,** 220. https://doi.org/10.1186/gb-2010-11-12-220 (2010).

16. Knowles, D. G., Roder, M., Merkel, A. & Guigo, R. Grape RNA-Seq analysis pipeline environment. *Bioinformatics* **29,** 614–621. https://doi.org/10.1093/bioinformatics/btt016 (Jan. 2013).

17. Abbas-Aghababazadeh, F., Li, Q. & Fridley, B. L. Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLOS ONE* **13** (ed Lin, H.) e0206312. https://doi.org/10.1371/journal.pone.0206312 (Oct. 2018).

18. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628. https://doi.org/10.1038/nmeth.1226 (May 2008).

19. Zhao, S., Ye, Z. & Stanton, R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* **26,** 903–909. https://doi.org/10.1261/rna.074922.120 (Apr. 2020).

20. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* **131,** 281–285. https://doi.org/10.1007/s12064-012-0162-3 (Aug. 2012).

21. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. *Genome Biology* **17.** https://doi.org/10.1186/s13059-016-0940-1 (Apr. 2016).

22. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17.** https://doi.org/10.1186/s13059-016-0881-8 (Jan. 2016).

23. Goh, W. W. B., Wang, W. & Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology* **35,** 498–507. https://doi.org/10.1016/j.tibtech.2017.02.012 (June 2017).

24. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127. https://doi.org/10.1093/biostatistics/kxj037 (Apr. 2006).

25. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics* **3** (ed Gibson, G.) e161. https://doi.org/10.1371/journal.pgen.0030161 (Sept. 2007).

26. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428. https://doi.org/10.1038/nature06758 (Mar. 2008).

27. Fatkin, D., Seidman, C. E. & Seidman, J. G. Genetics and Disease of Ventricular Muscle. *Cold Spring Harbor Perspectives in Medicine* **4,** a021063–a021063. https://doi.org/10.1101/cshperspect.a021063 (Jan. 2014).

28. Rodriguez-Esteban, R. & Jiang, X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Medical Genomics* **10.** https://doi.org/10.1186/s12920-017-0293-y (Oct. 2017).

29. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics* **47,** 569–576. https://doi.org/10.1038/ng.3259 (Apr. 2015).

30. Baldi, P. & Hatfield, G. W. *DNA microarrays and gene expression: from experiments to data analysis and modeling* (Cambridge university press, 2011).

31. Spellman, P. T. *et al.* Comprehensive Identification of Cell Cycle–regulated Genes of the YeastSaccharomyces cerevisiaeby Microarray Hybridization. *Molecular Biology of the Cell* **9** (ed Fink, G. R.) 3273–3297. https://doi.org/10.1091/mbc.9.12.3273 (Dec. 1998).

32. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752. https://doi.org/10.1038/35021093 (Aug. 2000).

33. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences* **98,** 13790–13795. https://doi.org/10.1073/pnas.191502998 (Nov. 2001).

34. Nachtomy, O., Shavit, A. & Yakhini, Z. Gene expression and the concept of the phenotype. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **38,** 238–254. https://doi.org/10.1016/j.shpsc.2006.12.014 (Mar. 2007).

35. Kandpal, R. P., Saviola, B. & Felton, J. The era of 'omics unlimited. *BioTechniques* **46,** 351–355. https://doi.org/10.2144/000113137 (Apr. 2009).

36. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21.** https://doi.org/10.1186/s13059-020-1926-6 (Feb. 2020).

37. Marx, V. The big challenges of big data. *Nature* **498,** 255–260. https://doi.org/10.1038/498255a (June 2013).

38. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41,** D991–D995. https://doi.org/10.1093/nar/gks1193 (Nov. 2012).

39. Athar, A. *et al.* ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Research* **47,** D711–D715. https://doi.org/10.1093/nar/gky964 (Oct. 2018).

40. *Sequence Read Archive* https://www.ncbi.nlm.nih.gov/sra. Accessed: 2020-11-25.

41. Ziemann, M., Eren, Y. & El-Osta, A. Gene name errors are widespread in the scientific literature. *Genome Biology* **17.** https://doi.org/10.1186/s13059-016-1044-7 (Aug. 2016).

42. Leonelli, S. *Data-centric biology: A philosophical study* (University of Chicago Press, 2016).

43. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47,** D766–D773. https://doi.org/10.1093/nar/gky955 (Oct. 2018).

44. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10, 000 Tumors from 33 Types of Cancer. *Cell* **173,** 291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022 (Apr. 2018).

45. Hunter, L. in *Artificial Intelligence and Molecular Biology* 1–46 (American Association for Artificial Intelligence, USA, 1993). ISBN: 0262581159.

46. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74. https://doi.org/10.1038/nature15393 (Sept. 2015).

47. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* ISBN: 9780387848846. https://books.google.es/books?id=eBSgoAEACAAJ (Springer, 2009).

48. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nature Methods* **15,** 399–400. https://doi.org/10.1038/s41592-018-0019-x (May 2018).

49. Lever, J., Krzywinski, M. & Altman, N. Model selection and overfitting. *Nature Methods* **13,** 703–704. https://doi.org/10.1038/nmeth.3968 (Aug. 2016).

50. Wang, Y. *et al.* Gene selection from microarray data for cancer classification—a machine learning approach. *Computational Biology and Chemistry* **29,** 37–46. https://doi.org/10.1016/j.compbiolchem.2004.11.001 (Feb. 2005).

51. Mramor, M., Leban, G., Demšar, J. & Zupan, B. *Conquering the Curse of Dimensionality in Gene Expression Cancer Diagnosis: Tough Problem, Simple Models* in *Artificial Intelligence in Medicine* (eds Miksch, S., Hunter, J. & Keravnou, E. T.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2005), 514–523. ISBN: 978-3-540-31884-2.

52. Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A. & Ploner, A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* **21,** 3017–3024. https://doi.org/10.1093/bioinformatics/bti448 (Apr. 2005).

53. Bland, J. M. & Altman, D. G. Statistics notes: Multiple significance tests: the Bonferroni method. *BMJ* **310,** 170–170. https://doi.org/10.1136/bmj.310.6973.170 (Jan. 1995).

54. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (Jan. 1995).

55. Kuhn, A. *et al.* Cell population-specific expression analysis of human cerebellum. *BMC Genomics* **13,** 610. https://doi.org/10.1186/1471-2164-13-610 (2012).

56. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35,** i436–i445. https://doi.org/10.1093/bioinformatics/btz363 (July 2019).

57. Cobos, F. A., Vandesompele, J., Mestdagh, P. & Preter, K. D. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34,** 1969–1979. https://doi.org/10.1093/bioinformatics/bty019 (Jan. 2018).

58. Repsilber, D. *et al.* Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* **11.** https://doi.org/10.1186/1471-2105-11-27 (Jan. 2010).

59. Gong, T. *et al.* Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PLoS ONE* **6** (ed Rattray, M.) e27156. https://doi.org/10.1371/journal.pone.0027156 (Nov. 2011).

60. *pracma: Practical Numerical Math Functions* https://cran.r-project.org/web/packages/pracma/index.html. Accessed: 03-09-2020.

61. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12,** 453–457. https://doi.org/10.1038/nmeth.3337 (Mar. 2015).

62. Dong, L., Kollipara, A., Darville, T., Zou, F. & Zheng, X. Semi-CAM: A semi-supervised deconvolution method for bulk transcriptomic data with partial marker gene information. *Scientific Reports* **10.** https://doi.org/10.1038/s41598-020-62330-2 (Mar. 2020).

63. Li, Y. & Xie, X. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics* **14.** https://doi.org/10.1186/1471-2105-14-s5-s11 (Apr. 2013).

64. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18.** https://doi.org/10.1186/s13059-017-1349-1 (Nov. 2017).

65. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462,** 108–112. https://doi.org/10.1038/nature08460 (Oct. 2009).

66. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6,** 377–382. https://doi.org/10.1038/nmeth.1315 (Apr. 2009).

67. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine* **50.** https://doi.org/10.1038/s12276-018-0071-8 (Aug. 2018).

68. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165,** 1012–1026. https://doi.org/10.1016/j.cell.2016.03.023 (May 2016).

69. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546,** 431–435. https://doi.org/10.1038/nature22794 (June 2017).

70. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications* **10.** https://doi.org/10.1038/s41467-018-08023-x (Jan. 2019).

71. Braga, F. A. V. *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nature Medicine* **25,** 1153–1163. https://doi.org/10.1038/s41591-019-0468-5 (June 2019).

72. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36,** 411–420. https://doi.org/10.1038/nbt.4096 (Apr. 2018).

73. Wu, J. T. *et al.* Behind the scenes: A medical natural language processing project. *International Journal of Medical Informatics* **112,** 68–73. https://doi.org/10.1016/j.ijmedinf.2017.12.003 (Apr. 2018).

74. Sheikhalishahi, S. *et al.* Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics* **7,** e12239. https://doi.org/10.2196/12239 (Apr. 2019).

75. Wei, W.-Q. *et al.* Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *Journal of the American Medical Informatics Association* **19,** 219–224. https://doi.org/10.1136/amiajnl-2011-000597 (Mar. 2012).

76. Liao, K. P. *et al.* Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic Disease Cohorts. *PLOS ONE* **10** (ed Bamias, G.) e0136651. https://doi.org/10.1371/journal.pone.0136651 (Aug. 2015).

77. Sarker, A. & Gonzalez, G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics* **53,** 196–207. https://doi.org/10.1016/j.jbi.2014.11.002 (Feb. 2015).

78. Manning, C., Manning, C., Schütze, H. & SCHUTZE, H. *Foundations of Statistical Natural Language Processing* ISBN: 9780262133609. https://books.google.es/books?id=ZL34DwAAQBAJ (MIT Press, 1999).

79. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* **5,** 135–146. ISSN: 2307-387X (2017).

80. Honnibal, M. & Montani, I. *spaCy* version 2.3.2. July 2020. https://github.com/explosion/spaCy.

81. Savova, G. K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* **17,** 507–513. https://doi.org/10.1136/jamia.2009.001560 (Sept. 2010).

82. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* **32,** 267D–270. https://doi.org/10.1093/nar/gkh061 (Jan. 2004).

83. Yetisgen-Yildiz, M. & Pratt, W. The Effect of Feature Representation on MEDLINE Document Classification. *AMIA Annu Symp Proc.,* 849–853 (2005).

84. Mescher, A. L. *Junqueira's Basic Histology Text and Atlas* 15th. ISBN: 978-1-26-002618-4 (McGraw-Hill Education, 2018).

85. He, L., Long, L. R., Antani, S. & Thoma, G. R. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine* **107,** 538–556. https://doi.org/10.1016/j.cmpb.2011.12.007 (Sept. 2012).

86. Gupta, E., Bhalla, P., Khurana, N. & Singh, T. Histopathology for the diagnosis of infectious diseases. *Indian Journal of Medical Microbiology* **27,** 100. https://doi.org/10.4103/0255-0857.49423 (2009).

87. Prewitt, J. M. S. & Mendelsohn, M. L. The analysis of cell images. *Annals of the New York Academy of Sciences* **128,** 1035–1053. https://doi.org/10.1111/j.1749-6632.1965.tb11715.x (Dec. 2006).

88. Pantanowitz, L., Farahani, N. & Parwani, A. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International,* 23. https://doi.org/10.2147/plmi.s59826 (June 2015).

89. Srinidhi, C. L., Ciga, O. & Martel, A. L. *Deep neural network models for computational histopathology: A survey* 2019. arXiv: 1912.12378 [eess.IV].

90. Mikula, S., Trotts, I., Stone, J. M. & Jones, E. G. Internet-enabled high-resolution brain mapping and virtual microscopy. *NeuroImage* **35,** 9–15. https://doi.org/10.1016/j.neuroimage.2006.11.053 (Mar. 2007).

91. Besson, S. *et al.* in *Digital Pathology* 3–10 (Springer International Publishing, 2019). https://doi.org/10.1007/978-3-030-23937-4_1.

92. Satyanarayanan, M., Goode, A., Gilbert, B., Harkes, J. & Jukic, D. OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics* **4,** 27. https://doi.org/10.4103/2153-3539.119005 (2013).

93. Gurcan, M. *et al.* Histopathological Image Analysis: A Review. *IEEE Reviews in Biomedical Engineering* **2,** 147–171. https://doi.org/10.1109/rbme.2009.2034865 (2009).

94. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* **42,** 60–88. https://doi.org/10.1016/j.media.2017.07.005 (Dec. 2017).

95. Sirinukunwattana, K. *et al. Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest* 2016. arXiv: 1603.00275 [cs.CV].

96. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369,** 1318–1330. https://doi.org/10.1126/science.aaz1776 (Sept. 2020).

97. Dimitriou, N., Arandjelović, O. & Caie, P. D. Deep Learning for Whole Slide Image Analysis: An Overview. *Frontiers in Medicine* **6.** https://doi.org/10.3389/fmed.2019.00264 (Nov. 2019).

98. Kong, B., Wang, X., Li, Z., Song, Q. & Zhang, S. in *Lecture Notes in Computer Science* 236–248 (Springer International Publishing, 2017). https://doi.org/10.1007/978-3-319-59050-9_19.

99. Momeni, A., Thibault, M. & Gevaert, O. Deep Recurrent Attention Models for Histopathological Image Analysis. https://doi.org/10.1101/438341 (Oct. 2018).

100. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* 2015. arXiv: 1505.04597 [cs.CV].

101. Quellec, G., Cazuguel, G., Cochener, B. & Lamard, M. Multiple-Instance Learning for Medical Image and Video Analysis. *IEEE Reviews in Biomedical Engineering* **10,** 213–234. https://doi.org/10.1109/rbme.2017.2651164 (2017).

102. Genome.gov. *The GTEx Project.* https://www.genome.gov/Funded-Programs-Projects/Genotype-Tissue-Expression-Project. Accessed: 01-06-2020.

103. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45,** 580–585. https://doi.org/10.1038/ng.2653 (May 2013).

104. *GTEx Portal* http://www.gtexportal.org. Accessed: 2020-11-25.

105. *Database of Genotypes and Phenotypes* https://www.ncbi.nlm.nih.gov/gap/. Accessed: 2020-11-25.

106. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348,** 660–665. https://doi.org/10.1126/science.aaa0355 (May 2015).

107. Biorepositories and Biospecimen Research Branch. *GTEx Standard Operating Procedures Library* Accessed:16-07-2020. 2015. https://biospecimens.cancer.gov/resources/sops/library.asp.

108. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2,** 559–572. https://doi.org/10.1080/14786440109462720 (Nov. 1901).

109. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24,** 417–441. https://doi.org/10.1037/h0071325 (1933).

110. Abraham, G. & Inouye, M. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLoS ONE* **9** (ed Zhang, Y.) e93766. https://doi.org/10.1371/journal.pone.0093766 (Apr. 2014).

111. Yeung, K. Y. & Ruzzo, W. L. Principal component analysis for clustering gene expression data. *Bioinformatics* **17,** 763–774. https://doi.org/10.1093/bioinformatics/17.9.763 (Sept. 2001).

112. Tsuyuzaki, K., Sato, H., Sato, K. & Nikaido, I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biology* **21.** https://doi.org/10.1186/s13059-019-1900-3 (Jan. 2020).

113. Jolliffe, I. & Springer-Verlag. *Principal Component Analysis* ISBN: 9780387954424 (Springer, 2002).

114. Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7,** 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x (Sept. 1936).

115. Turk, M. & Pentland, A. *Face recognition using eigenfaces* in *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition* (1991), 586–587.

116. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html (2008).

117. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nature Communications* **10.** https://doi.org/10.1038/s41467-019-13056-x (Nov. 2019).

118. Hinton, G. E. & Roweis, S. T. in *Advances in Neural Information Processing Systems 15* (eds Becker, S., Thrun, S. & Obermayer, K.) 857–864 (MIT Press, 2003). http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding.pdf.

119. Van der Maaten, L. & Hinton, G. Visualizing non-metric similarities in multiple maps. *Machine Learning* **87,** 33–55. https://doi.org/10.1007/s10994-011-5273-4 (Dec. 2011).

120. Wattenberg, M., Viégas, F. & Johnson, I. How to Use t-SNE Effectively. *Distill.* http://distill.pub/2016/misread-tsne (2016).

121. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* **37,** 38–44. https://doi.org/10.1038/nbt.4314 (Dec. 2018).

122. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* 2018. arXiv: 1802.03426 [stat.ML].

123. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R* ISBN: 1461471370 (Springer Publishing Company, Incorporated, 2014).

124. Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of Machine Learning* ISBN: 026201825X (The MIT Press, 2012).

125. Wolpert, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* **8,** 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341 (Oct. 1996).

126. Romagnoni, A., Jégou, S., Steen, K. V., Wainrib, G. & Hugot, J.-P. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Scientific Reports* **9.** https://doi.org/10.1038/s41598-019-46649-z (July 2019).

127. Ke, G. *et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree* in *Advances in Neural Information Processing Systems 30 (NIP 2017)* (2017). https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/.

128. Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system* in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), 785–794.

129. Claesen, M. & Moor, B. D. *Hyperparameter Search in Machine Learning* 2015. arXiv: 1502.02127 [cs.LG].

130. Snoek, J., Larochelle, H. & Adams, R. P. *Practical Bayesian Optimization of Machine Learning Algorithms* 2012. arXiv: 1206.2944 [stat.ML].

131. Jones, D. R., Schonlau, M. & Welch, W. J. *Journal of Global Optimization* **13,** 455–492. https://doi.org/10.1023/a:1008306431147 (1998).

132. Adadi, A. & Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **6,** 52138–52160. https://doi.org/10.1109/access.2018.2870052 (2018).

133. Gilpin, L. H. *et al. Explaining Explanations: An Overview of Interpretability of Machine Learning* 2018. arXiv: 1806.00069 [cs.AI].

134. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1,** 206–215. https://doi.org/10.1038/s42256-019-0048-x (May 2019).

135. Moretti, S. *et al.* Combining Shapley value and statistics to the analysis of gene expression data in children exposed to air pollution. *BMC Bioinformatics* **9.** https://doi.org/10.1186/1471-2105-9-361 (Sept. 2008).

136. Zuallaert, J. *et al.* SpliceRover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics* **34** (ed Hancock, J.) 4180–4188. https://doi.org/10.1093/bioinformatics/bty497 (June 2018).

137. Shapley, L., Roth, A. & Press, C. U. *The Shapley value: essays in honor of Lloyd S. Shapley* ISBN: 9780521361774. https://books.google.es/books?id=JK7MKu2A9cIC (Cambridge University Press, 1988).

138. Lipovetsky, S. & Conklin, M. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry* **17,** 319–330. https://doi.org/10.1002/asmb.446 (2001).

139. Lundberg, S. M. & Lee, S.-I. in *Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al.*) 4765–4774 (Curran Associates, Inc., 2017). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

140. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* **41,** 647–665. https://doi.org/10.1007/s10115-013-0679-x (Aug. 2013).

141. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2,** 2522–5839 (2020).

142. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* **2,** 749–760. https://doi.org/10.1038/s41551-018-0304-0 (Oct. 2018).

143. Pan-cancer analysis of whole genomes. *Nature* **578,** 82–93. https://doi.org/10.1038/s41586-020-1969-6 (Feb. 2020).

144. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org (MIT Press, 2016).

145. Nwankpa, C., Ijomah, W., Gachagan, A. & Marshall, S. Activation Functions: Comparison of trends in Practice and Research for Deep Learning. *CoRR* **abs/1811.03378.** arXiv: 1811.03378. http://arxiv.org/abs/1811.03378 (2018).

146. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536. https://doi.org/10.1038/323533a0 (Oct. 1986).

147. Sun, R. *Optimization for deep learning: theory and algorithms* 2019. arXiv: 1912.08957 [cs.LG].

148. Hanin, B. & Rolnick, D. *How to Start Training: The Effect of Initialization and Architecture* 2018. arXiv: 1803.01719 [stat.ML].

149. Masters, D. & Luschi, C. Revisiting Small Batch Training for Deep Neural Networks. *CoRR* **abs/1804.07612.** arXiv: 1804.07612. http://arxiv.org/abs/1804.07612 (2018).

150. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* **15,** 20170387. https://doi.org/10.1098/rsif.2017.0387 (Apr. 2018).

151. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444. https://doi.org/10.1038/nature14539 (May 2015).

152. LeCun, Y. *et al.* Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* **1,** 541–551 (1989).

153. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* **36,** 193–202. https://doi.org/10.1007/bf00344251 (Apr. 1980).

154. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324 (1998).

155. Dhillon, A. & Verma, G. K. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence* **9,** 85–112. https://doi.org/10.1007/s13748-019-00203-0 (Dec. 2019).

156. Khan, A., Sohail, A., Zahoora, U. & Qureshi, A. S. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *CoRR* **abs/1901.06032.** arXiv: 1901.06032. http://arxiv.org/abs/1901.06032 (2019).

157. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Molecular Systems Biology* **12,** 878. https://doi.org/10.15252/msb.20156651 (July 2016).

158. Hirsch, F. R. *et al.* The Prognostic and Predictive Role of Histology in Advanced Non-small Cell Lung Cancer: A Literature Review. *Journal of Thoracic Oncology* **3,** 1468–1481. https://doi.org/10.1097/jto.0b013e318189f551 (Dec. 2008).

159. Smoller, B. R. Histologic criteria for diagnosing primary cutaneous malignant melanoma. *Modern Pathology* **19,** S34–S40. https://doi.org/10.1038/modpathol.3800508 (Jan. 2006).

160. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine* **25,** 1054–1056. https://doi.org/10.1038/s41591-019-0462-y (June 2019).

161. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* **1,** 800–810. https://doi.org/10.1038/s43018-020-0085-8 (July 2020).

162. Schmauch, B. *et al.* Transcriptomic learning for digital pathology. https://doi.org/10.1101/760173 (Sept. 2019).

163. Ash, J. T., Darnell, G., Munro, D. & Engelhardt, B. E. Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. https://doi.org/10.1101/458711 (Oct. 2018).

164. Chen, R. J. *et al. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis* 2020. arXiv: 1912.08937 [cs.CV].

165. Sing, T. *et al.* A deep learning-based model of normal histology. https://doi.org/10.1101/838417 (Nov. 2019).

166. Gerke, S., Minssen, T. & Cohen, G. in *Artificial Intelligence in Healthcare* 295–336 (Elsevier, 2020). https://doi.org/10.1016/b978-0-12-818438-7.00012-5.

167. Stanta, G. & Bonin, S. Overview on Clinical Relevance of Intra-Tumor Heterogeneity. *Frontiers in Medicine* **5.** https://doi.org/10.3389/fmed.2018.00085 (Apr. 2018).

168. Asp, M., Bergenstråhle, J. & Lundeberg, J. Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. *BioEssays* **42,** 1900221. https://doi.org/10.1002/bies.201900221 (May 2020).

169. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82. https://doi.org/10.1126/science.aaf2403 (June 2016).

170. Maniatis, S. *et al.* Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364,** 89–93. https://doi.org/10.1126/science.aav9776 (Apr. 2019).

171. Satija Lab. *stxBrain.SeuratData: 10X Genomics Visium Mouse Brain Dataset* R package version 0.1.1 (2019).

172. *Analysis, visualization, and integration of spatial datasets with Seurat.* https://satijalab.org/seurat/v3.2/spatial_vignette.html. Accessed: 2020-11-27.

173. Rosso, J. Q. D. & Levin, J. The Clinical Relevance of Maintaining the Functional Integrity of the Stratum Corneum in both Healthy and Disease-affected Skin. *J Clin Aesthet Dermatol* **4,** 22–42 (2011).

174. Kumar, G. & Bhatia, P. K. *A Detailed Review of Feature Extraction in Image Processing Systems* in *2014 Fourth International Conference on Advanced Computing & Communication Technologies* (IEEE, Feb. 2014). https://doi.org/10.1109/acct.2014.74.

175. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *CoRR* **abs/1912.01703.** arXiv: 1912.01703. http://arxiv.org/abs/1912.01703 (2019).

176. Abadi, M. *et al. TensorFlow: A system for large-scale machine learning* in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (2016), 265–283. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

177. Deng, J. *et al. ImageNet: A Large-Scale Hierarchical Image Database* in *CVPR09* (2009).

178. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big Data* **3.** https://doi.org/10.1186/s40537-016-0043-6 (May 2016).

179. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *CoRR* **abs/1411.1792.** arXiv: 1411.1792. http://arxiv.org/abs/1411.1792 (2014).

180. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* **6.** https://doi.org/10.1186/s40537-019-0197-0 (July 2019).

181. *Torchvision models* https://pytorch.org/docs/stable/torchvision/models.html. Accessed: 2020-11-23.

182. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *CoRR* **abs/1512.03385.** arXiv: 1512.03385. http://arxiv.org/abs/1512.03385 (2015).

183. Simonyan, K. & Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition* 2015. arXiv: 1409.1556 [cs.CV].

184. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. *Densely Connected Convolutional Networks* 2018. arXiv: 1608.06993 [cs.CV].

185. Li, S. *et al. PyTorch Distributed: Experiences on Accelerating Data Parallel Training* 2020. arXiv: 2006.15704 [cs.DC].

186. Howard, J. & Gugger, S. Fastai: A Layered API for Deep Learning. *Information* **11,** 108. ISSN: 2078-2489. http://dx.doi.org/10.3390/info11020108 (Feb. 2020).

187. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* 2017. arXiv: 1412.6980 [cs.LG].

188. Smith, L. N. *Cyclical Learning Rates for Training Neural Networks* 2017. arXiv: 1506.01186 [cs.CV].

189. *General Structure of the Digestive System* https://training.seer.cancer.gov/anatomy/digestive/structure.html. Accessed: 2020-11-25.

190. Kokhlikyan, N. *et al. Captum: A unified and generic model interpretability library for PyTorch* 2020. arXiv: 2009.07896 [cs.LG].

191. Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic Attribution for Deep Networks* 2017. arXiv: 1703.01365 [cs.LG].

192. Dhamdhere, K., Sundararajan, M. & Yan, Q. How Important Is a Neuron? *CoRR* **abs/1805.12233.** arXiv: 1805.12233. http://arxiv.org/abs/1805.12233 (2018).

193. Ilse, M., Tomczak, J. M. & Welling, M. *Attention-based Deep Multiple Instance Learning* 2018. arXiv: 1802.04712 [cs.LG].

194. Lu, M. Y. *et al. Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images* 2020. arXiv: 2004.09666 [eess.IV].

195. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20,** 273–282. https://doi.org/10.1038/s41576-018-0088-9 (Jan. 2019).

196. Madissoon, E. *et al.* scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biology* **21.** https://doi.org/10.1186/s13059-019-1906-x (Dec. 2019).

197. Bacon, E. R., Ihle, K., Lee, P. P. & Waisman, J. R. Building a rapid autopsy program – a step-by-step logistics guide. *Translational Medicine Communications* **5.** https://doi.org/10.1186/s41231-020-00074-x (Nov. 2020).

198. Jiang, L. *et al.* A Quantitative Proteome Map of the Human Body. *Cell* **183,** 269–283.e19. https://doi.org/10.1016/j.cell.2020.08.036 (Oct. 2020).

199. Simonovsky, E., Schuster, R. & Yeger-Lotem, E. Large-scale analysis of human gene expression variability associates highly variable drug targets with lower drug effectiveness and safety. *Bioinformatics* **35** (ed Valencia, A.) 3028–3037. https://doi.org/10.1093/bioinformatics/btz023 (Jan. 2019).

200. Zhang, K. *et al.* Evaluating the effects of causes of death on postmortem interval estimation by ATR-FTIR spectroscopy. *International Journal of Legal Medicine* **134,** 565–574. https://doi.org/10.1007/s00414-019-02042-z (Mar. 2019).

201. Locci, E. *et al.* A 1H NMR metabolomic approach for the estimation of the time since death using aqueous humour: an animal model. *Metabolomics* **15.** https://doi.org/10.1007/s11306-019-1533-2 (May 2019).

202. Liu, R. *et al.* Predicting postmortem interval based on microbial community sequences and machine learning algorithms. *Environmental Microbiology* **22,** 2273–2291. https://doi.org/10.1111/1462-2920.15000 (Apr. 2020).

203. Javan, G. T. *et al.* Identification of cadaveric liver tissues using thanatotranscriptome biomarkers. *Scientific Reports* **10.** https://doi.org/10.1038/s41598-020-63727-9 (Apr. 2020).

204. Gallins, P., Saghapour, E. & Zhou, Y.-H. Exploring the Limits of Combined Image/'omics Analysis for Non-cancer Histological Phenotypes. *Frontiers in Genetics* **11.** https://doi.org/10.3389/fgene.2020.555886 (Oct. 2020).

205. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6.** https://doi.org/10.7554/elife.27041 (Dec. 2017).

206. Liu, T., Nogueira, L. B., Lleo, A. & Conesa, A. Transcriptional differences for COVID-19 Disease Map genes between males and females indicate a different basal immunophenotype relevant to the disease. https://doi.org/10.1101/2020.09.30.321059 (Oct. 2020).

207. Rastelli, D. *et al.* Androgen regulation of bowel function in mice and humans. https://doi.org/10.1101/2020.10.15.341081 (Oct. 2020).

208. Goodman, W. A., Erkkila, I. P. & Pizarro, T. T. Sex matters: impact on pathogenesis, presentation and treatment of inflammatory bowel disease. *Nature Reviews Gastroenterology & Hepatology* **17,** 740–754. https://doi.org/10.1038/s41575-020-0354-0 (Sept. 2020).

209. Hoffman, G. E. *et al.* Sex differences in the human brain transcriptome of cases with schizophrenia. https://doi.org/10.1101/2020.10.05.326405 (Oct. 2020).

210. Allen, A. M., Scheuermann, T. S., Nollen, N., Hatsukami, D. & Ahluwalia, J. S. Gender Differences in Smoking Behavior and Dependence Motives Among Daily and Nondaily Smokers. *Nicotine & Tobacco Research* **18,** 1408–1413. https://doi.org/10.1093/ntr/ntv138 (June 2015).

211. Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *The Lancet Oncology* **20,** e253–e261. https://doi.org/10.1016/s1470-2045(19)30154-8 (May 2019).

212. Raab., S. S., Nakhleh, R. E. & Ruby, S. G. Patient Safety in Anatomic Pathology: Measuring Discrepancy Frequencies and Causes. *Archives of Pathology and Laboratory Medicine* **129,** 459–466. ISSN: 0003-9985 (Apr. 2005).

213. Tellez, D. *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis* **58,** 101544. https://doi.org/10.1016/j.media.2019.101544 (Dec. 2019).

214. Tong, L., Sha, Y. & Wang, M. D. *Improving Classification of Breast Cancer by Utilizing the Image Pyramids of Whole-Slide Imaging and Multi-scale Convolutional Neural Networks* in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)* **1** (2019), 696–703.

215. Sureka, M., Patil, A., Anand, D. & Sethi, A. *Visualization for Histopathology Images using Graph Convolutional Neural Networks* 2020. arXiv: 2006.09464 [eess.IV].

216. Takahashi, T. *et al.* Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature.* https://doi.org/10.1038/s41586-020-2700-3 (Aug. 2020).

217. Haslbauer, J. D. *et al.* Retrospective Post-mortem SARS-CoV-2 RT-PCR of Autopsies with COVID-19-Suggestive Pathology Supports the Absence of Lethal Community Spread in Basel, Switzerland, before February 2020. *Pathobiology,* 1–11. https://doi.org/10.1159/000512563 (Nov. 2020).

218. *Artificial Intelligence: Shaping Europe's digital future* https://ec.europa.eu/digital-single-market/en/artificial-intelligence. Accessed: 2020-12-02.

219. Brown, T. B. *et al. Language Models are Few-Shot Learners* 2020. arXiv: 2005.14165 [cs.CL].

220. LeNail, A. NN-SVG: Publication-Ready Neural Network Architecture Schematics. *Journal of Open Source Software* **4,** 747. https://doi.org/10.21105/joss.00747 (Jan. 2019).